Multi-modal Open World User Identification

BAHAR IRFAN, Centre for Robotics and Neural Systems, University of Plymouth, United Kingdom
MICHAEL GARCIA ORTIZ, AI Lab, SoftBank Robotics Europe, France and City, University of London, United Kingdom
NATALIA LYUBOVA, Prophesee, France
TONY BELPAEME, IDLab - imec, Ghent University, Belgium and Centre for Robotics and Neural Systems, University of Plymouth, United Kingdom

User identification is an essential step in creating a personalised long-term interaction with robots. This requires learning the users continuously and incrementally, possibly starting from a state without any known user. In this paper, we describe a multi-modal incremental Bayesian network with online learning, which is the first method that can be applied in such scenarios. Face recognition is used as the primary biometric, and it is combined with ancillary information, such as gender, age, height and time of interaction, to improve the recognition. The Multi-modal Long-term User Recognition Dataset is generated to simulate various human-robot interaction (HRI) scenarios and evaluate our approach in comparison to face recognition, soft biometrics and a state-of-the-art open world recognition method (Extreme Value Machine). The results show that the proposed methods significantly outperform the baselines, with an increase in the identification rate up to 47.9% in open-set and closed-set scenarios, and a significant decrease in long-term recognition performance loss. The proposed models generalise well to new users, provide stability, improve over time, and decrease the bias of face recognition. The models were applied in HRI studies for user recognition, personalised rehabilitation and customer-oriented service, which showed that they are suitable for long-term HRI in the real world.

CCS Concepts: • **Mathematics of computing → Bayesian networks**; • **Security and privacy → Biometrics**; • **Theory of computation → Online learning algorithms**; • **Computer systems organization → Robotics**; • **Information systems → Personalization**.

Additional Key Words and Phrases: Open world recognition, Bayesian network, soft biometrics, incremental learning, online learning, multi-modal dataset, long-term user recognition, Human-Robot Interaction

ACM Reference Format:
Bahar Irfan, Michael Garcia Ortiz, Natalia Lyubova, and Tony Belpaeme. 2021. Multi-modal Open World User Identification. *ACM Trans. Hum.-Robot Interact.* 11, 1, Article 6 (October 2021), 51 pages. https://doi.org/10.1145/3477963

Authors' addresses: Bahar Irfan, bahar.irfan@plymouth.ac.uk, Centre for Robotics and Neural Systems, University of Plymouth, Drake Circus, Plymouth, United Kingdom, PL48AA; Michael Garcia Ortiz, AI Lab, SoftBank Robotics Europe, 43 rue du Colonel Pierre Avia, Paris, France, 75015 and City, University of London, Northampton Square, London, United Kingdom, EC1V 0HB, michael.garcia-ortiz@city.ac.uk; Natalia Lyubova, Prophesee, 74 Rue du Faubourg Saint-Antoine, Paris, France, 75012, nlyubova@prophesee.ai; Tony Belpaeme, IDLab - imec, Ghent University, Technologiepark-Zwijnaarde 126, Ghent, Belgium, 9052 and Centre for Robotics and Neural Systems, University of Plymouth, Drake Circus, Plymouth, United Kingdom, PL48AA, tony.belpaeme@ugent.be.

2573-9522/2021/10-ART6
https://doi.org/10.1145/3477963

1 INTRODUCTION

User identification is an important step towards achieving and maintaining a personalised long-term interaction with robots. For instance, a user may need to be identified for providing personalised rehabilitation therapy [41]. When a robot is first deployed, it will start from a "tabula rasa" state with no prior knowledge of users. As users are encountered over a possibly extended period of time, their identity and information are stored by the robot. Hence, the system has to identify enrolled and "unknown" users, which is known as *open-set identification*. Open-set identification is a well-established field [48, 76, 77], but in a real-world setting, these unknown users might need to be added into the system for future recognition. One solution is to retrain the system after introducing a novel user. However, this requires storing the previous samples, which could create a prohibitively large computational burden in long-term deployments. Furthermore, it would require a significant amount of time to retrain with a growing number of users and samples [8]. Instead, the system should allow scaling and support incremental learning of new classes, which is termed *open world recognition* [8].

Face recognition (FR), i.e., identifying a person based on their face, has been the most prominent technique in biometric identification due to its non-intrusive character. Most state-of-the-art methods use deep learning based approaches [68, 79–81], but only a few approaches exist for open-set recognition [9, 33]. Most models are not suitable for open world recognition due to the *catastrophic forgetting* problem, which refers to the drastic loss of performance on previously learned classes when a new class is introduced [62, 63, 66]. Existing approaches that could help to overcome this problem often require a part of the previous data for retraining, which might not be available.

Incremental learning is not sufficient for adapting to changes in the environment. For instance, an algorithm designed for open world recognition may not be able to recognise a person after a new haircut, because the model is not updated for known samples. Humans show a good model for recognition because they can continuously adapt to changing circumstances by updating their prior beliefs, known as *online learning*, and use multi-modal information instead of a single biometric

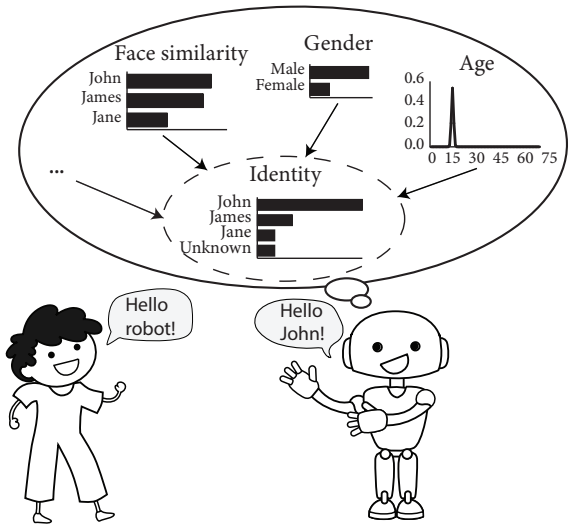


Fig. 1. Robots can make use of multi-modal information to recognise users more accurately in long-term interactions.

(modality) for estimation of the identity, such as recognising a person from the voice in a dark room. Biometric systems that combine multiple biometric traits or attributes obtained through the same sensor (e.g., face and iris [16, 21, 83, 87]) or various sensors (e.g., face and voice [10, 17, 18, 58, 82]) for establishing identity are known as *multi-modal biometric systems* [24, 47]. Most robots are also suitable for multi-modal recognition, as they have multiple sensors and perception algorithms (as shown in Fig. 1), which allow them to recognise users even when data are inaccurate or noisy, for example, in the case of image blur or illumination changes [85]. Moreover, the combination of multi-modal data can help overcome issues related to similarities between users[1], by differentiating on additional available information, for example, age and gender. Such ancillary physical or behavioural characteristics, called *soft biometrics*, can be used to improve the recognition performance [24, 45, 47]. Combining multi-modal recognition with online learning can improve recognition further in time. For instance, a user can be initially mistaken for another in certain circumstances, but these variations can be learned over time and combined with other modalities to improve recognition where FR fails.

In our earlier work [43], we proposed a multi-modal weighted Bayesian Network with online learning, which is the first approach for combining soft biometrics (gender, age, height and time of interaction) with a primary biometric (face recognition) for open world user identification in real-time human-robot interaction (HRI). This model, here referred to as Multi-modal Incremental Bayesian Network (MMIBN), is the first method for sequential and incremental learning in open world user recognition that allows starting from a state without any known users (i.e., it does not require preliminary training to recognise users and it can learn new users incrementally). This work showed that the proposed model is suitable for real-world human-robot interaction experiments for user recognition in real-time. However, the limited population size (14 users) and the narrow age range (24-40) of the users in that experiment prevented us from claiming that the results can be generalised for application in larger populations. On the other hand, obtaining a dataset that encapsulates a diverse set of characteristics for a large number of users over long-term interactions is a laborious task in HRI. Thus, we created the Multi-modal Long-Term User Recognition Dataset[2], which contains images of 200 users (with age range 10 to 63) with name, gender, age and height labels, along with artificially generated height estimations and various time of interactions to simulate a long-term HRI scenario. We obtained the images from the largest publicly available dataset of face images with gender and age labels, IMDB-WIKI dataset[3] [71, 72]. To obtain the multi-modal biometric information from these images (face, gender and age estimations), we used (NAOqi) proprietary algorithms of the Pepper robot[4], similar to our earlier work.

Our main contribution is the extension of our earlier work [43] to take in multi-modal information, typically available in HRI, to markedly increase user identification and subsequently improve user experience in long-term interactions for a large number of users in a variety of settings. We also provide a detailed description of the Multi-modal Incremental Bayesian Network, highlighting the mathematical formulations and assumptions behind the models that were not addressed in [43]. In addition, we present our findings from applying the optimised models in long-term HRI experiments in the real world [41–43]. Correspondingly, we make the following contributions (source code, multi-modal dataset, trained models and results on the dataset are available online[2]):

[1]https://www.wired.com/story/10-year-old-face-id-unlocks-mothers-iphone-x/

[2]Latest version of the Multi-modal Incremental Bayesian Network: https://github.com/birfan/MultimodalRecognition Multi-modal Long-Term User Recognition Dataset, source code used in this work and the corresponding results and the trained models are available at: https://github.com/birfan/MultimodalRecognitionDataset

[3]https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

[4]https://www.softbankrobotics.com/corp/robots/

- creating the Multi-modal Long-Term User Recognition Dataset with 200 users of varying characteristics
- introducing *long-term recognition performance loss*
- combining optimal normalisation methods for each parameter in the Bayesian network in a *hybrid* approach
- formulating the proposed online learning in terms of Expectation Maximization (EM) and Maximum Likelihood (ML)
- applying Bayesian optimisation on the weights of the soft biometric identifiers and the quality of the estimation
- evaluating the proposed model against a state-of-the-art open world recognition method (Extreme Value Machine [73])
- evaluating the stability of the model for learning users sequentially (similar to batch learning) and at random intervals (similar to a real-world scenario)
- evaluating the generalisability of the model for new users (performance during training set in comparison to open-set and closed-set recognition)
- evaluating the model for varying frequency of user appearances (modelled with uniform and Gaussian timing of interaction, and varying dataset sizes)
- evaluating the progress of the model over time (with the increasing number of recognitions)
- analysing recognition bias in face recognition, the proposed approach and Extreme Value Machine
- evaluating the models on the data from the real-world HRI study (4 weeks) in [43] in comparison to the corresponding optimised models
- evaluating the model in a real-world (5-day) HRI study with a personalised barista robot at an international student campus in Paris (France)
- evaluating the models in a long-term (5-months) HRI study within a cardiac rehabilitation programme at a hospital in Bogotá (Colombia)

The rest of the paper is organised as follows: Section 2 gives a brief overview of the current practice of open world recognition, online learning, multi-modal biometrics algorithms, and user recognition in human-robot interaction (HRI). Section 3 describes the methodology and the structure of the proposed Bayesian network. Section 4 describes the recognition module for NAOqi that is used to obtain the multi-modal biometric information for the proposed model. Section 5 explains the procedure of the creation of the Multi-modal Long-Term User Recognition Dataset. Section 6 presents the empirical evaluation of the proposed methods on closed-set and open-set datasets. Section 7 highlights the implications of the results and discusses the initial assumptions. Section 8 evaluates the optimised models in long-term HRI studies in the real world. Section 9 concludes with a summary of the work.

2 RELATED WORK

Our work lies at the intersection of open world recognition, online learning, multi-modal biometrics, and HRI.

2.1 Open World Recognition

One of the first algorithms applied to open world recognition was Nearest-Non Outlier (NNO) [8], which modified Nearest Class Mean (NCM) [64] for open-set classification and incremental learning. Another approach is Extreme Value Machine (EVM) [73] based on Extreme Value Theory, which outperformed NNO on the open world ImageNet benchmark [8, 73]. However, both of these methods work with incrementally adding a batch of new classes (e.g., 100 at a time), as opposed

to incremental learning of classes (one at a time). Similarly, the approach proposed in [29] is based on a center-based similarity space learning method and 1-vs-rest strategy of Support Vector Machines (SVM) for object classification. However, none of these methods has been evaluated on user recognition.

2.2 Online Learning

Several online learning methods exist for various application areas [34]. In video-based recognition, Lee and Kriegman [55] proposed an online learning algorithm of probabilistic appearances, but a prior generic model is necessary for this approach. Boucenna et al. [13] used online and incremental learning in two neural networks for facial expression recognition and face/non-face discrimination in an HRI imitation game. The former neural network uses a k-means variant SAW (Self Adaptive Winner takes all) [49] to categorise focus points in the image, whereas the latter predicts the interaction rhythm [3] (i.e., timing for interaction) to detect whether the user is interacting with the robot. While the face discrimination method was shown to generalise to new users successfully, the facial expression recognition achieved low success rates for generalisation. In addition, both approaches required preliminary training, and were evaluated on a low number of users (20). De Rosa et al. [26] used online learning in open world (object) recognition for incremental learning of a classification metric, the threshold for novelty detection and describing the space of classes. The approach was applied to three existing algorithms, namely, NCM, NNO and Nearest Ball Classifier (NBC) [27]. Their results showed that online learning increases classification performance.

2.3 Multi-Modal Biometrics

In a multi-modal biometric system, information from different identifiers, such as face recognition or gender identification, is fused via prior or post classification [44]. Prior classification requires access to the features or sensor values of the identifiers, which are generally not available for proprietary algorithms. For post-classification, two approaches exist: *classification* and *combination* of confidence scores. Classification methods, such as neural networks and SVM, combine non-homogeneous data from individual classifiers into a feature vector for further classification without the need for preprocessing. In the combination approach, individual matching scores from the identifiers are combined into a scalar score in three steps: (1) normalisation of scores into a common domain, (2) combination of scores based on Bayes decision rule and posterior probabilities, e.g., *sum* or *product rule*, and (3) thresholding for classification. The performance of these approaches depends on the chosen method and threshold.

Bayesian approaches have been widely used for combining primary biometrics, such as face and speaker recognition [10, 17, 82], as well as combining soft biometrics [25, 45, 46, 67, 78, 86]. For instance, Jain et al. [45] proposed a Bayesian network for combining fingerprints with soft biometric traits, namely, gender, ethnicity, and height. They used a fixed weighting scheme, where the biometrics with smaller variability and more substantial distinguishing capability were given more weight and achieved slight improvement in recognition. Similarly, Scheirer et al. [78] used a Bayesian network with Noisy-OR weighting that combines face recognition with ethnicity, hair colour, gender, age, eyebrow type and non-soft biometric contextual information, such as the occupation and location of the person. Contrary to the work in [45] and our approach, they used the accuracy of estimators to adjust the FR match score.

2.4 User Recognition in Human-Robot Interaction

Similar to biometric recognition, the most common approach for user recognition in HRI is through FR [4, 5, 23, 32, 38]. However, robots can take advantage of multi-modal recognition due to the variety of different sensors they carry. Soft biometrics are especially important because they allow

non-intrusive recognition, but only a few studies use soft biometrics. Martinson et al. [61] used a weighted summation of soft biometrics (clothing, complexion and height) to identify users within a short-term interaction from a group of only three users. Boucenna et al. [12] gathered extensive data (100 images per person) during a game and later evaluated the recognition offline using a Hebbian rule-based neural network. Ouellet et al. [65] combined face recognition, speaker identification, and human metrology through Hampel estimators in closed-set identification using a substantial time for training (3.5 minutes) and a small number of participants (pretraining on 22, test on 7). Al-Qaderi and Rad [1] combined face, body and speech information using a spiking neural network in closed-set identification and have evaluated on a simulated dataset. These approaches do not apply to open world recognition, hence, their methods are not easily comparable to ours.

Our previous work [43] introduced a multi-modal weighted Bayesian network, which is the first approach in combining multi-modal biometric information for sequential and incremental learning of new users for open world recognition that allows starting from a state without any known users. It is also the first approach in combining soft biometrics (gender, age, height and time of interaction) with a primary biometric (FR) to identify a user in real-time HRI. Online learning was used for learning the likelihoods of the network from sequential data to improve the recognition over long-term interactions. The weights of the network were optimised to minimise the number of incorrect recognitions. The *quality of the estimation* measure was introduced to decrease the number of incorrect recognitions for unknown users. The results obtained in a user study with 14 participants over four weeks showed a slight improvement in identification rate (up to 1.4% in open-set and 4.4% in closed-set recognition) compared to 90.3% of FR. The optimised weights suggested that age is the least effective soft biometric parameter, whereas height is the most effective one. Moreover, the Bayesian network performed worse with online learning. However, we concluded that the dataset might be biased towards the participants' characteristics due to the low number of participants and limited age range, and an evaluation with a bigger dataset is necessary to understand the capabilities of the system entirely.

This paper extends the work in [43], for evaluating the approach within the Multi-modal Long-Term User Recognition Dataset and two other real-world HRI experiments, and optimising the weights of the Bayesian network through a *long-term recognition performance loss* criterion with *hybrid* normalisation.

3 MULTI-MODAL INCREMENTAL BAYESIAN NETWORK

A Bayesian network is a probabilistic graphical model which represents conditional dependencies of a set of variables through a directed acyclic graph. Bayesian networks are suitable for combining scores of identifiers with uncertainties when the knowledge of the world is incomplete [78].

We developed a weighted multi-modal incremental Bayesian network (MMIBN), integrating multi-modal biometric information for reliable recognition in open world identification through a

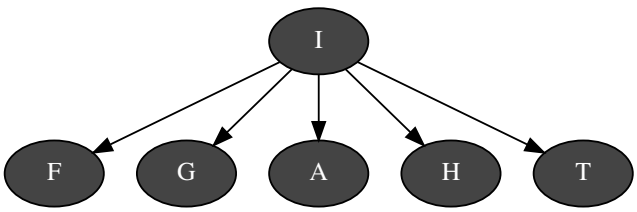


Fig. 2. The naive Bayesian network model with identity (I), face (F), gender (G), age (A), height (H), and time of interaction (T) nodes.

naive Bayes model (see Fig. 2). The naive Bayes classifier model assumes conditional independence between predictors, which is a reasonable assumption for a multi-modal biometric identifier as the individual identifiers do not affect each other's results. The architecture for the estimation of the user identity (I) in MMIBN and the recognition process are presented in Fig. 14 and 15 in Appendix A. The primary biometric in our system is face recognition (F), which is fused with soft biometrics, namely, gender (G), age (A), and height (H) estimations, in addition to the time of interaction (T), which can be distinguishing if the users are encountered at patterned interaction times, such as for weekly appointments in rehabilitation. We hypothesise that the integration of these soft biometrics will reduce the effects of noisy data, as described in Section 1, and increase the identification rate. Nonetheless, the MMIBN allows extension with other primary biometric traits, such as voice and fingerprint, and other soft biometrics, such as eye colour and gait, to improve recognition. The pyAgrum[5] [36] library is used for implementing the Bayesian network structure. Parts of MMIBN were previously described in our prior work [43], however, this section provides the underlying mathematical formulations and full details of the system for reproducibility, and introduces the long-term recognition performance loss (Section 3.6) and hybrid normalisation (Section 3.7).

3.1 Structure

The number of states for each node depends on the modality: F and I nodes have n_e+1 states, where n_e is the number of enrolled (known) users. A and H nodes are restricted to the available range of the identifier, such as $[0, 75]$ for A and $[50, 240]$ (cm) for H. G has "female" and "male" states. T is defined by the day of the week and the time, through time *slots*. For example, if each minute corresponds to a time slot (i.e., time period, t_p, is 1 min), there will be 10080 T states (there are 10080 minutes in a week).

When a user is encountered, the corresponding multi-modal biometric evidence is collected from the identifiers. An example for the biometric evidence from the identifiers and the transformed (weighted and normalised) evidence is shown in Fig. 16B in Appendix A. FR provides similarity scores, which give the percentage of similarity of the user to the known faces in the database. Age, height, and time are assumed to be discrete random variables with a discretised and normalised normal distribution of probabilities, $N(\mu, \sigma^2)$, defined by (1), where V is the estimated value, Z is the standard score, and C is the confidence of the biometric indicator for the estimated value.

$$\mu = V, \quad P(\frac{-0.5}{\sigma} < Z < \frac{0.5}{\sigma}) = C \tag{1}$$

The time period and its standard deviation (σ_t in the normal distribution of T) can be set depending on the precision required in the application. A smaller time period and standard deviation ensure higher precision, however, this would increase the complexity of the Bayesian network, thereby increasing the time to identify the user. In addition, a higher precision carries the risks of decreasing the recognition rate, if the users are not encountered near the time slot that they were previously seen. For example, if users in the application scenario will change every 5 minutes, then $t_p = 5$ min and $\sigma_t = 15$ min would be reasonable. On the other hand, in an HRI scenario, $t_p = 30$ min with $\sigma_t = 60$ min can allow better identification because it is less likely to encounter users around the same time every day. Hence, we use the latter in this paper.

3.2 Weights of the Network

Soft biometric traits are characteristics that are not suited to identify an individual uniquely. We can assume that the population will have similar characteristics, but the distribution is unknown. However, some soft biometric features may contain more information about an individual than

[5]https://agrum.gitlab.io/

others, e.g., age is often more informative than gender. This can be modelled by using different weights for the parameters in a Bayesian network [45].

Weights (w_i) are used as the exponential to the likelihoods of the child nodes (X_i), similar to the work in [88]. In contrast to our previous work [43], we optimise the weights of soft biometric features (gender, age, height and time of interaction) through Bayesian optimisation, as described in Appendix C.6, while the weight of the face node (w_F) is set to be 1, as it is the only primary biometric in our system. The posterior probability $P(I^j|X_1,..,X_n)$ is approximated as in (2). I^j stands for the jth user ($I = j$), where I is the identity node.

$$P(I^j|X_1,..,X_n) \propto \frac{P(I^j)\prod_i P(X_i|I^j)^{w_i}}{P(X_1,..,X_n)} \tag{2}$$

As in [45], we assume that the identifiers perform equally well on all users. Therefore, the accuracy of an identifier is independent of the user and equal priors are assumed for each of the identifiers. The posterior probability simplifies to the equation shown in (3).

$$P(I^j|X_1,..,X_n) \propto P(I^j)\prod_i P(X_i|I^j)^{w_i} \tag{3}$$

Because the distribution of users over time is not known, one approach for determining $P(I^j)$ is to use adaptive priors using frequencies of user appearance, however, this can create a bias in the system towards the most frequently observed user as it affects the posterior probability directly, thus, may result in a decrease in the identification rate. Therefore, we assume that the probability of encountering user j is equally likely as encountering user m, hence, we assume equal priors for $P(I)$, as shown in (4), where n_e is the number of enrolled users, which is updated whenever a new user is enrolled, as presented in Fig. 16 in Appendix A.

$$P(I^j) = P(I) = \frac{1}{n_e} \tag{4}$$

3.3 Quality of the Estimation

Algorithms for open-set problems generally use a threshold (e.g., over the highest probability/score) to determine if the user is already enrolled or "unknown". However, the resulting posterior probabilities in a Bayesian network can be low due to the multiplication of the conditionally independent modalities and vary depending on the number of states. Hence, we use the two-step ad hoc mechanism introduced in [43] to transform the Bayesian network to allow open-set recognition: (1) An "Unknown" (U) state is used in both F and I nodes. The similarity score in FR of U is set to the FR threshold (θ_{FR}), such that when normalised, scores below/above the threshold will have lower/higher probabilities than U. This allows maintaining the threshold for the FR system in use. (2) We use the confidence measure called the *quality of the estimation* (Q). Given the evidence y_t at time t, it compares the highest posterior probability (P_w) to the second highest (P_s), as shown in (5). The difference between the probabilities decreases, as the number of enrolled users (n_e) increases since $\sum_j P(I^j|y_t) = 1.0$. A similar method was used in [31] for estimating the quality of localisation based on different images.

$$Q = [P_w(I^j|y_t) - P_s(I^j|y_t)] * n_e \tag{5}$$

Using the quality of the estimation enables decreasing misidentifications. For example, the highest posterior score can be very high, but if the second highest posterior is very close to it, then it means that there are two possible strong candidates for the current user. If the system were to identify the user in this case, the resulting misidentification could cause adverse effects

on the current user especially in the case of different genders or age differences between the two users, as well as security issues. Thus, it is more preferable to identify the user as unknown, if the quality is zero or below a predetermined threshold (θ_Q), or if U has the highest posterior probability. Otherwise, the identity is estimated with a maximum a posteriori (MAP) estimation, given in (6).

$$j^* = \begin{cases} U, & \text{if } Q = 0 \text{ or } Q < \theta_Q \text{ or} \\ & P(I^U|y_t) > P(I^j|y_t) \text{ for all } j \\ \arg\max_j P(I^j|y_t), & \text{otherwise} \end{cases} \tag{6}$$

3.4 Incremental Learning

For personalisation in long-term HRI applications, new users may often need to be enrolled in a system to allow recognition in subsequent encounters, such as for admitting a new patient to personalised robot therapy. However, in such applications, the intermediary (e.g., clinical staff) and end users (e.g., patients) are often non-experts, hence, systems that require the least amount of technical knowledge, effort and time are desirable, especially those that allow users to enrol themselves. Thus, we developed an incremental learning system for the weighted multi-modal Bayesian network, which expands the network upon new user enrolment. When the MMIBN detects that the user is new, the robot requests to meet the user, and (verbally) asks for their name, gender, birth year, and height, which the user can enter through a tablet interface, after which a photo of the user is taken by the robot (step 9 in Fig. 15). This information, along with the time of interaction, is gathered to have the ground truth values for recognition, and for setting the initial likelihoods of the MMIBN.

Initially, the system starts from a "tabula rasa" state, where there are no known users. Bayesian network is formed when the first user is enrolled: one state for the new user and one for the "Unknown" (U) state. Fig. 16A (in Appendix A) illustrates an example for the initial MMIBN after the enrolment of the first user, e.g., a 25-years-old female who is 168 cm tall and encountered at 11:00 am on a Monday. The initial likelihood for F is set to be much higher for the true values as shown in (7), where w_F is the weight of the face variable, and n_e is the number of enrolled users. The value was found based on preliminary experiments.

$$P(F^k|I^j) = \begin{cases} 0.9^{w_F}, & \text{if } k = j \\ [0.1/(n_e - 1)]^{w_F}, & \text{otherwise} \end{cases} \tag{7}$$

The remaining likelihoods are set using the prior knowledge that the user entered in a similar structure to the evidence for age, height and time variables with a discretised and normalised normal distribution, $N(\mu, \sigma^2)$, where μ is the true value (e.g., age of the person), and σ is the standard deviation of the identifier. Gender is set at $[0.99^{w_G}, 0.01^{w_G}]$ ratio, which is experimentally found. For the unknown state, $P(X_i^k|I^U)$ is set to be uniformly distributed, as an unknown user can be of any age, height and be recognised at any time of the day, except for the face node, which follows (7).

When a new user is enrolled, the Bayesian network is expanded by adding a new state to I and F nodes. $P(F^k|I^j)$ for each previous state in I (including U) is updated by appending the value corresponding to $k \neq j$ condition in (7), and then probabilities are re-normalised. The likelihoods of G, A, H and T nodes for the previously enrolled users remain the same. An example of the MMIBN likelihoods during incremental learning of a new user, e.g., a 37-years-old male, 173 cm tall, and encountered on a Wednesday at 8:00 pm, is illustrated in Fig. 16E in Appendix A.

The scalability feature removes the need to retrain the network when a new user is introduced, hence, the time complexity is decreased, which can be crucial if the new user is introduced at

a later step (e.g., after 1000 users). More precisely, if each image corresponding to $\overline{n_o}$ average number of observations per user was to be recognised again after a new user is added to the face database, it would take a significant amount of time to expand the network compared to scaling, since $n_e * \overline{n_o} * O(FR) \gg n_e * O(1)$ updates, where $O(FR)$ is the time complexity of the FR algorithm, and n_e is the number of enrolled users.

In order to reduce the risk of confusing new users with known users, it is preferable to have sufficient data within the MMIBN prior to making reliable estimations, hence, in the first few recognitions (here, we chose $N < N_{min} = 5$ recognitions, i.e., the first 4 recognitions)[6], the identity is declared as unknown, regardless of the estimated identity, as illustrated in Fig. 16C (Appendix A).

3.5 Online Learning of Likelihoods

Bayesian network parameters are generally determined by expert opinion or by learning from data [51]. The former can cause incorrect estimations if the set probabilities are not accurate enough. The latter, for which Maximum Likelihood (ML) estimation is commonly used, is not possible when the Bayesian network is constructed with incomplete data. One solution is to use offline batch learning, however, it requires storing data that can cause memory problems in long-term interactions. Another approach is to update the parameters as the data arrive, which is termed online learning. Variants of Expectation Maximization (EM) algorithm with a learning rate (EM(η)) [6, 20, 57, 59] have been proposed for online learning in Bayesian networks.

We use a Bayesian network where the likelihoods are updated through EM(η) with an adaptive η (learning rate) based on ML estimation, similar to Voting EM [20]. Adopting the notation in [6], the formulation is given in (8). θ^t_{ijk} represents the likelihood of the modality X_i at time t, $P(X_i = x^k_i | I^j)$. $P_{\theta^t}(x^k_i | y_t, I^j)$ represents the posterior probability of the modality X_i at time t given the current evidence y_t and the actual identity of the user I^j. The difference between Voting EM and our approach is that we work with continuous probabilities due to uncertainties in the identifiers. We will refer to the proposed multi-modal incremental Bayesian network with online learning as MMIBN:OL.

$$\theta^{t+1}_{ijk} = \begin{cases} \eta_j P_{\theta^t}(x^k_i | y_t, I^j) + (1 - \eta_j)\theta^t_{ijk}, & \text{if } P(I^j) = 1 \\ \theta^t_{ijk}, & \text{otherwise} \end{cases} \tag{8}$$

Combining ML estimate to achieve an adaptive learning rate (given in (9)) allows the learning rate to depend on the observation of the user j (n_{oj}), which is more reliable than using a fixed rate for all users. Also, each observation of the user creates a progressively smaller update on the likelihoods, such that, the effect of a new observation decreases as the number of recognitions of the user increases.

$$\eta_j = \frac{1}{n_{oj} + 1} \tag{9}$$

Supervised learning is necessary to achieve accurate online learning. The identity of the user should be known for updating the corresponding likelihoods, which can be achieved in HRI by asking for a confirmation of the estimated identity.

If the user j is previously enrolled in the system, the likelihoods are only updated for user j, as shown in Fig. 16D (in Appendix A) based on the evidence in Fig. 16B. On the other hand, if

[6]This parameter can be set to another value (including 0) in the algorithm for MMIBN. Increasing this value would allow the MMIBN to produce more reliable estimations of new users, however, this could also decrease the identification performance of known users. Hence, we chose a sufficiently low value. It is also possible to use the identity estimated by face recognition (instead of declaring unknown identity) in the algorithm for the first few recognitions.

the user j is a new user, online learning is applied on the face likelihood for the unknown state ($P(F^k|I^U)$), followed by incremental learning by expanding the MMIBN (as described in Section 3.4), and finally by applying online learning for the new user, as illustrated in steps 8-18 in Fig. 15 and in Fig. 16F. The likelihoods of gender, age, height, and time remain the same for U to ensure uniform distribution.

3.6 Long-Term Recognition Performance Loss

The standard metrics for open-set identification are Detection and Identification Rate (DIR) and False Alarm Rate (FAR) [69]. DIR is the fraction of correctly classified probes (samples) within the probes of the enrolled users ($\mathscr{P}_{\mathscr{E}}$), given in (10). FAR is the fraction of incorrectly classified probes within the probes of unknown users ($\mathscr{P}_{\mathscr{U}}$), given in (11).

$$DIR = \frac{|\{\arg\max_j P(I^j|y_t) = j | j, j \in \mathscr{P}_{\mathscr{E}}\}|}{|\mathscr{P}_{\mathscr{E}}|} \tag{10}$$

$$FAR = \frac{|\{\arg\max_j P(I^j|y_t) = j | k, j \in \mathscr{P}_{\mathscr{E}}, k \in \mathscr{P}_{\mathscr{U}}\}|}{|\mathscr{P}_{\mathscr{U}}|} \tag{11}$$

In other words, DIR represents the "true positive" (TP) of enrolled users, in which the current probe (referring to the multi-modal biometric sample) belongs to a previously enrolled user and identified correctly. FAR serves as a "false positive" (FP) for unknown users, that is, the probe belongs to an unknown user, but he/she is identified as an enrolled user. However, TP and FP are notions of *verification* problems, in which the probe is compared against a claimed identity, thus, are generally not applicable to *open-set identification*. Instead, the trade-off between DIR and FAR that depends on the threshold of the identifier, is generally represented by a Receiver Operating Characteristic (ROC) curve. The standard practice in biometric identification is to determine the desired FAR, which would then set the threshold and DIR.

Depending on the biometric application, the cost of incorrectly identifying a user as known may be very different from the cost of incorrect identification of the enrolled user [47]. For short-term interactions, in which a user will be encountered $1-2$ times, FAR is as important or more important than DIR. However, for long-term interactions, users will be encountered a greater number of times. Thus, correctly identifying a user (in a closed-set) becomes more important than correctly identifying an unknown user (open-set). Hence, we introduce the *long-term recognition performance loss* (L) that creates a balance between DIR and FAR based on the average number of observations per user ($\overline{n_o}$), as presented in (12), where α is the ratio of importance of DIR compared to FAR.

Weights of MMIBN are optimised through this loss function, for gender, age, height and time in $[0, 1]$ range, along with quality (Q) that can change within $[0, 0.5]$ range. Ideally $L = 0$, where all unknown users are identified as such (FAR= 0.0) and the known users are correctly identified (DIR= 1.0).

$$\begin{aligned} L &= \alpha * (1 - DIR) + (1 - \alpha) * FAR \\ \alpha &= 1 - \frac{1}{\overline{n_o}} \end{aligned} \tag{12}$$

3.7 Normalisation Methods

The scores from each modality must be normalised into a common range (e.g., $[0, 1]$) to ensure a meaningful combination. It is important to choose a method that is insensitive to outliers and provides a good estimate of the distribution [44], such as, minmax, tanh [37], softmax [11], and normsum (dividing each value by the sum of values). We introduce *hybrid* normalisation which

combines the methods that achieve the lowest loss for each modality. In other words, hybrid normalisation uses the best performing normalisation method for each modality. Extensive tests were made on the dataset obtained from our previous work in [43] to get the optimal methods for each modality (*F*, *G*, *A*, *H* and *T*). The long-term recognition performance loss was compared for each combination of the individual modality with face recognition (*F*, *F-G*, *F-A*, *F-H*, *F-T*) by optimising the weights for each of the combinations. The resulting hybrid normalisation uses normsum for face, gender, and height; tanh for age; softmax for time of interaction.

4 RECOGNITION MODULE

While MMIBN can be applied on other platforms, its main purpose is for enabling incremental user recognition in long-term human-robot interaction in the real world. The proposed approach does not require heavy computing, therefore, it is suitable for use on commercially available robots. We employ this system on Pepper and NAO[7] robots, which are amongst the most commonly used robots in HRI research [53], for our experiments (as described in Section 8). These robots are operated by NAOqi[8] software, which includes different modules that allowed us to extract face similarity scores, gender, height and age estimations from a single image through the Recognition Module in Fig. 13 (Appendix A). The internal states of the proprietary algorithms (developed by OKAO) are inaccessible, hence, we assume that the gender and age estimations are not used to obtain the face similarity scores, and they are conditionally independent of the FR results, even though they are obtained from the 2D image. The height estimation in NAOqi is measured through the 3D sensor (in the eyes) of the Pepper robot, and based on the face position in the 2D image and the geometric transformations (based on the camera relative to the robot) for the NAO robot. Due to relying on only one primary biometric, in the absence of facial information, the user is not recognised since soft biometric information would not be sufficient to estimate the identity.

MMIBN can be used with any identifier software. The reason NAOqi identifiers are chosen is their capability for incremental recognition and their real-time performance, in other words, these algorithms work on a single CPU on a robot without requiring preliminary training. In contrast, the state-of-the-art deep learning methods for face recognition (such as Dlib [50]) are not optimised for low computational power systems, hence, they may require a vast amount of time for encoding images, recognition and retraining[9], which makes them unsuitable for real-time open world user recognition on a robot. Similarly, OpenFace[10] [2], which is an implementation of FaceNet [79] and a popular closed-set face identification method, was found to be unsuitable for real-world HRI, because the classifier needs to be retrained after a new user enrolment with all the available data (instead of incremental learning) with batch learning of images for the new user, and the training time (albeit small) increases with the increasing number of users [2]. In addition, preliminary evaluations of OpenFace[11] provided unpromising results in new user identification. For instance,

[7] https://www.softbankrobotics.com/corp/robots/

[8] http://doc.aldebaran.com/2-5

[9] An implementation of Dlib for open world recognition using retraining on a dataset with a small number of users is explained in this link, which shows that a single recognition can take 6-7 seconds on a single CPU system: https://www.pyimagesearch.com/2018/06/18/face-recognition-with-opencv-python-and-deep-learning/

[10] https://cmusatyalab.github.io/openface

[11] The classifier demos in https://cmusatyalab.github.io/openface/demo-1-web/ and https://cmusatyalab.github.io/openface/demo-3-classifier/ were combined and applied on the Pepper robot. An image was taken from the robot's camera, identified with the pre-trained OpenFace celebrity classifier available at the latter link, and the confidence score of the classifier was displayed on the Pepper's tablet, along with the image of the user and the most similar celebrity. The confidence score of the classification ranges from 0% (user does not resemble any user in the database) to 100% (user is identical to a user in the database). If the confidence score is below (or equal to) 50%, the user is identified as unknown (new), as defined in the script for the former demo.

the first author was recognised consistently as Anne Hathaway with a high confidence (85 to 99.2%), despite the fact that the classifier was trained on only 10 users with 600 images per user (i.e., the classifier must be very accurate in identifying known users), and the author does not resemble her that highly. Nevertheless, it is possible to use OpenFace or other identifiers, instead of the NAOqi user recognition algorithms, for obtaining the multi-modal biometric information for MMIBN.

5 MULTI-MODAL LONG-TERM USER RECOGNITION DATASET

Our prior work provided evidence that the proposed model is suitable for long-term HRI in the real world. However, the optimised parameters of the model could not be generalised to a larger population due to the limited number of users and their narrow age range in that study. On the other hand, collecting a diverse training set within a long-term real-world HRI scenario is very challenging. To the best of our knowledge, the only publicly available dataset that contains the soft biometrics used in our system (except for the time of interaction) with a dataset of faces is BioSoft [74]. However, due to the low number of subjects (75), and the lack of numeric height values, we decided to create our own Multi-modal Long-Term User Recognition Dataset.

Datasets that contain images in the form of "mugshots", such as NIST Mugshot Identification Database[12], do not represent real-world HRI interactions in which the obtained images from the robot's camera may vary greatly depending on the users' actions and the environmental conditions. Therefore, it is important to use an image dataset with real-world variations, along with ground truth values of identity, gender and age of users to assess the performance of our model and the corresponding identifiers in similar conditions. The largest publicly available dataset of face images with gender and age labels is the IMDB-WIKI dataset [71, 72], which contains more than 500k images of 20k celebrities with a wide age range. As can be observed in Fig. 3, the images in this dataset may contain bad lighting conditions, occlusions, oblique viewing angles, a variety of facial expressions, partial faces of other people, face paint and disguise, and black and white images, because the images come from movies, TV series and events.

In addition to images, the estimated height of the user and the time of interaction with the robot would be necessary for user recognition in various HRI scenarios, where the users will be encountered sequentially over time. Thus, we created the Multi-modal Long-Term User Recognition

Fig. 3. Samples of images from the IMDB-WIKI dataset [71, 72] that are used in creating the Multi-modal Long-Term User Recognition Dataset.

[12]https://www.nist.gov/srd/nist-special-database-18

Dataset by (1) sampling a subset of the IMDB-WIKI image dataset, and (2) artificially generating height estimations and various time of interactions to simulate repeated encounters of the users with the robot. The resulting dataset contains 200 users (101 females, 98 males, and one transgender person, the age range is 10 to 63) with 10 to 41 images per user that adds up to 5735 images, height estimations and various (patterned and random) time of interactions, along with a database of users' names, genders, ages, and heights. Moreover, NAOqi identifier estimations (face similarity scores, gender and age estimations) are obtained for each image, and provided alongside the artificial height estimations and the time of interaction in order to simulate the information that would be acquired from a robot (e.g., NAO or Pepper) in an HRI scenario. The Multi-modal Long-Term User Recognition Dataset is available online[13].

5.1 Image Sampling

In the scope of this work, only one user is assumed to be present in each image, hence, the cropped faces of IMDB dataset is used. To simulate an open world HRI scenario, where the users will be met in consecutive days or weeks, we chose images of users that are from the same year. Furthermore, we assume that the average number of times a user will be observed is $\overline{n_o} \geq 10$, which is a reasonable assumption for long-term HRI. Hence, we choose celebrities who have more than 10 images each corresponding to the same age. Moreover, to assess the incremental learning capabilities of our model with a user database that is more realistic for HRI (i.e., sufficiently large with 100 to 200 users instead of thousands of users), we (randomly) sampled 200 users out of 20k celebrities.

In order to create a diverse set of ages in the dataset, the images that correspond to an age that is within the five most common ages (25, 26, 28, 30, 31) in the set were randomly rejected (with 50% probability) during the selection. For instance, Anne Hathaway has sufficient images corresponding to 25 and 27 years old in the IMDB-WIKI dataset. However, 25 is among the five most common ages, thus, with a 50% chance, this set of images were excluded from the selection, hence, the images of Anne Hathaway corresponding to 27 years old were chosen instead. This also resulted in some celebrities who only have images corresponding to a certain age in the dataset to be excluded from the selection. The resulting age range is $10 - 63$, with the mean age of 33.04 (SD= 9.28).

Subsequently, the dataset is cleaned in three steps: by removing (1) images with a resolution lower than 150x150, (2) images without a face detected by NAOqi, (3) images that erroneously correspond to another person. Furthermore, in order of user appearance (as detailed further in Section 6.2), NAOqi identifiers are applied on the selected images to obtain face similarity scores, gender and age estimations. If the user has not been previously encountered, the same image is used to identify the user before and after enrolment to the face database in NAOqi.

5.2 Height and Time of Interaction

Height was found to be the most important soft biometric in determining the identity in [43]. To validate whether this finding persists for a large number of users with diverse characteristics, and optimise its weight for applying it to real-world HRI experiments, we artificially created height data for each user. To keep the data realistic and model the differences between the estimated heights, Gaussian noise with $\sigma = 6.3$ cm (as found in [43] for NAOqi height estimation) is added to the actual heights of the users obtained from the web.

Given our assumption that the users will be encountered at least 10 times in long-term HRI, we created two datasets: (1) D-Ten, where each user is observed precisely ten times, e.g., ten return visits to a robot therapist, and (2) D-All, in which each user is encountered a different amount of times (10 to 41 times). Two types of distribution are considered for the time of interaction: (1)

[13] https://github.com/birfan/MultimodalRecognitionDataset

patterned interaction times in a week modelled through a Gaussian mixture model, where the user will be encountered certain times on specific days, which applies to HRI in rehabilitation and education areas, and (2) random interaction times represented by uniform distribution, such as in domestic applications with companion robots, where the user can be seen at any time of the day in the week. As a result, we created four (sub)datasets as part of the Multi-modal Long-Term User Recognition Dataset: D-Ten$_{\text{Uniform}}$, D-Ten$_{\text{Gaussian}}$, D-All$_{\text{Uniform}}$, D-All$_{\text{Gaussian}}$.

6 EVALUATION

In this section, we evaluate our proposed models based on the hypotheses presented in Section 6.1. The procedure of creating the cross-validation sets is described in Section 6.2. Initially, the parameters of the multi-modal incremental Bayesian network (Section 6.3) are optimised for open world recognition in long-term interactions in Section 6.4. Using those parameters, the model is compared to face recognition and soft biometrics on the Multi-modal Long-Term User Recognition Dataset for the training set, closed-sets and open-set tests in Section 6.5.

6.1 Hypotheses

H1 Our proposed multi-modal incremental Bayesian network will improve user recognition compared to face recognition alone. As measured by a decrease in the long-term recognition performance loss (L) and an increase in the identification rate of known users (DIR).

H2 Online learning will improve user recognition over a non-adaptive model. As measured by a decrease in L and an increase in DIR.

H3 Hybrid normalisation will outperform the individual normalisation methods.

H4 When assumptions are made about the temporal interaction pattern of the user, recognition will improve. When the time of interaction is uniformly distributed, the loss L will be higher.

These hypotheses will be validated with various analyses, as provided in Table 1.

Table 1. The analyses for validating the hypotheses and the corresponding results. A check mark represents a support for the hypothesis, a cross mark represents rejecting the hypothesis, and the crossed check mark represents partial support for the hypothesis.

Analysis	Section	H1	H2	H3	H4
Normalisation methods	Appendix C.5		✗	✓	✓✗
Tukey's HSD on loss	Section 6.5.1	✓	✗		✓
Tukey's HSD on DIR	Section 6.5.2	✓			✓
User identification in HRI	Section 8.1	✓	✗		
Barista robot	Section 8.2	✓			
Socially assistive robot	Section 8.3	✓✗	✓		

6.2 Procedure

Repeated k-fold cross-validation is used to evaluate the model stability and performance. The procedure is described in Algorithm 1 in Appendix B. Two methods for creating validation folds are used, namely, OrderedKFold and ShuffledKFold. OrderedKFold is the case where users are introduced one by one to the system without any repetitions of previous users during the enrolment. The order of repeated interactions is random after the enrolment. In ShuffledKFold, there can be repetitions of the previous user(s) before another user is introduced, because the order of overall samples is random. OrderedKFold is similar to batch learning in an incremental learning sense, whereas, the

iteration (repeat) created by ShuffledKFold is more similar to a real-world scenario. Our aim is to evaluate if there are any performance differences between the two cases and to prove that the model is stable across several repeats. A stratified random bin order is used for having a different initial bin and final bin in each fold to ensure a different enrolment order of users and a different test set, respectively. We chose K= 5 folds and R= 11 repeats.

Each dataset (D-Ten and D-All) is divided into two with 100 users each. The first set is then divided through cross-validation procedure with 80 − 20% ratio of data to the *training set* (first four bins, corresponding to 800 samples in D-Ten and 2308 in D-All) and *closed-set (training)* (final bin, corresponding to 200 samples in D-Ten, 578 in D-All). The *open-set* is created from the remaining 100 users (800 samples in D-Ten, 2280 in D-All). The *closed-set (open)* is similar to the *closed-set (training)*, which corresponds to the final bin in each fold (200 in D-Ten, 569 in D-All). The open-set evaluation is made by introducing the open-set samples after the training set, that is, 100 users are enrolled in the system, and recognised multiple times before the introduction of 100 new users. However, the results for the open-set do not include the results for training.

The only difference between Gaussian and uniform datasets is the time of the interaction for each sample; that is, the order of the samples is the same.

For online learning, the likelihoods are learned during the training phase (*training* and *open-set* cases), and the learned likelihoods are used without online learning for the *closed-set* cases.

6.3 Description of Variables

Given our datasets and the parameters of our model, we have four independent variables and three dependent variables for analysing the results on the evaluation sets: training, open-set, closed-set (training), closed-set (open). The dependent variables are DIR in (10), FAR in (11) and long-term recognition performance loss (shortly, loss) in (12). The independent variables are as follows:

(1) **Dataset size**: ten samples per user (D-Ten), random amount of samples (D-All)
(2) **Timing of interaction**: patterned interaction times (Gaussian), random interaction times (uniform)
(3) **Model**: non-adaptive MMIBN, MMIBN with online learning (MMIBN:OL)
(4) **Normalisation method**: softmax, minmax, tanh, normsum, and hybrid

6.4 Optimisation of Parameters

The parameters of the MMIBN need to be optimised to achieve the best recognition results. Correspondingly, we conducted several evaluations on the Multi-modal Long-Term User Recognition Dataset as described in detail in Appendix C. Here we summarise our findings for reasons of perspicuity.

Initially, the loss parameter α is set as 0.9, based on our average number of observations assumption ($\overline{n_o} = 10$) for long-term interaction (Appendix C.1). Subsequently, the optimum face recognition threshold with the lowest loss for (NAOqi) FR is found to be 0.4 (Appendix C.2).

MMIBN relies on the assumption that the multi-modal biometric information (face, gender, age, height and time of interaction) are conditionally independent given the identity of the user, since the individual identifiers do not affect each other's results. Accordingly, we assumed that the NAOqi identifiers (face, gender and age) are conditionally independent of each other, despite relying on the same visual input (2D image). Structural learning of the Bayesian network on the Multi-modal Long-Term User Recognition Dataset (in Appendix C.3) confirmed this assumption, showing that the naive Bayes classifier model is sufficient and suitable for multi-modal user identification, even when the modalities use the same input. Moreover, the average learned likelihoods in online learning are very close to the initially assumed network parameters in Section 3.4.

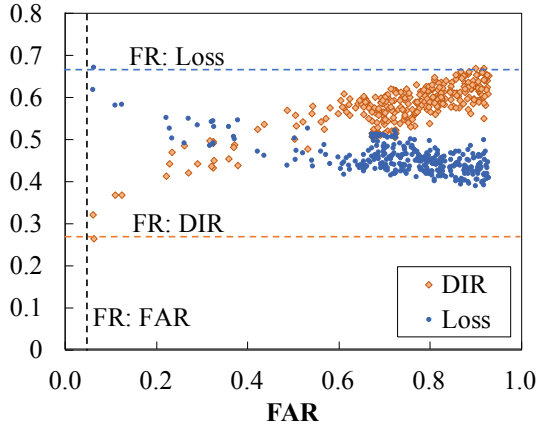


Fig. 4. ROC curve for MMIBN with hybrid normalisation in the all samples dataset with Gaussian times (D-All$_{Gaussian}$), with long-term recognition performance loss (Equation 12, represented with blue dots) for varying known user identification rate (DIR, represented with orange diamond shapes) and incorrect new user detection rate (FAR, x axis), for Bayesian optimisation of the weights and the quality of the estimation for 303 iterations over 5-fold cross-validation. Face recognition (FR) values are given in dashed lines (orange line representing DIR of FR, blue line for loss of FR, and black line for FAR of FR) for comparison. While optimising parameters to reduce the loss, DIR increases at the cost of increasing FAR.

Bayesian optimisation[14] is applied with these parameters to minimise the loss for each combination of the independent variables (40 conditions) by optimising the weights for soft biometrics and the threshold for the quality of the estimation (see Appendix C.6). Fig. 4 shows how the loss decreases during the optimisation, which results in an increase in DIR at the cost of an in increase FAR. The resulting loss of MMIBN is much lower than that of FR, and correspondingly DIR and FAR are much higher. Note that α can be adjusted to give more importance to FAR or a FAR can be set prior to optimisation, which may lead to a different set of optimised parameters.

While the average standard deviation of NAOqi age estimation is found to be higher (11.0) than in [43] (9.3), the age is found to be the most important parameter and height the least (see Appendix C.6), in contrast with the findings in [43]. Due to the higher number of users (200) and the diverse age range (10-63) in the Multi-modal Long-Term User Recognition Dataset, these results are more generalisable than our prior work. Moreover, when the ground truths are not taken into account, the standard deviation of age within the estimations is found to be 8.2, which is less than the average. This is due to the appearance of users (e.g., a 30-year-old person may look like 25), which suggests that online learning of likelihoods (MMIBN:OL) may provide better recognition performance over time, as the identifiers will get better at identifying users based on their own estimations instead of ground truth values. In addition, NAOqi gender recognition is found to be equally accurate for males and females with 0.9 as the recognition rate (i.e., users' genders are correctly recognised 90% of the time). Furthermore, using the confidence of the estimations instead of exclusively the estimated biometric data (e.g., estimated gender or age, as described in Section 3.1) allows overcoming deviations in the estimations.

With these optimised parameters, 11 repeats of 5-fold cross-validation were applied for each of the conditions (Appendix C.4), which showed that MMIBN models are stable across repeats (i.e., no significant difference in loss between repeats), and the models perform equally well for

[14]https://thuijskens.github.io/2016/12/29/bayesian-optimisation/

learning new users incrementally sequentially (OrderedKFold, similar to batch learning) and at random intervals (ShuffledKFold, similar to a real-world scenario). On the other hand, the size of the dataset, timing of interaction and normalisation method are found to have significant effects on the performance of the model, however, the non-adaptive model and the model with online learning performed equally well.

Hybrid normalisation is found to outperform the other normalisation methods in all conditions (Appendix C.5), supporting our hypothesis **H3**. The models achieved lower loss in D-All than in D-Ten, which showed that the proposed model gets better with the increasing number of recognitions. However, hybrid normalisation with online learning (MMIBN:OL) is found to perform worse than the non-adaptive model (MMIBN), in contrast with our hypothesis **H2**. Moreover, most methods are found to perform significantly worse when there is no interaction pattern (uniform timing of interaction), as compared to patterned (Gaussian) interactions, supporting our hypothesis **H4**.

6.5 Comparison to Baselines

On the grounds that the optimised parameters of our proposed MMIBN are found, we can compare its results to face recognition (FR) and soft biometrics (SB). FR results are obtained from the NAOqi estimations by setting FR threshold (θ_{FR}) to 0.4. SB results are obtained by giving zero weight to FR, that is, only gender and age estimates from NAOqi, artificial height estimates and time of interaction are used for identifying a user. The weights of these modalities in SB are the same as MMIBN, as shown in Fig. 19 (Appendix C.6). Similarly, the weights of SB:OL are the same as those of MMIBN:OL.

We transformed a state-of-the-art open world recognition method, Extreme Value Machine[15] [73] (EVM) to accept sequential and incremental data for online learning by adjusting its hyperparameters to use it as a baseline, as described in Appendix D. In the original work, batch learning of 50 classes were used with an average of 63806 data points at each update, instead of a single data point that we used in this work. We compared our methods with the performance of two EVM models: (a) EVM:FR, using NAOqi face recognition similarity scores as data, (b) EVM:MM, using multi-modal information in the same format as it is used for our methods.

Section 6.5.1 compares the long-term recognition performance loss (shortly, loss) between the models. Appendix C.4 provides evidence that there is a significant correlation between loss and DIR, and loss and FAR, but no significant correlation is found between DIR and FAR. Hence, the analysis of loss is sufficient to determine how the model performs in comparison to others. Nevertheless, we will report the results of FAR and DIR of the models in Section 6.5.2 to further observe how the open-set recognition metrics are affected.

6.5.1 Long-term Recognition Performance Loss. As previously mentioned, the proposed models perform better in terms of loss in D-All than in D-Ten, however, the results for D-Ten datasets show similar patterns to that of D-All. Taken the same number of recognitions for both D-All and D-Ten, that is equal to the number of samples in D-Ten for all evaluation sets, ANOVA shows that there is no significant difference in the sample size ($p = .67$) as the models perform equally well for D-All and D-Ten for the same number of samples. In other words, it does not matter if each user is observed the same number of times or not. This also supports that a higher number of samples increases the performance of the models. Hence, the following analysis will only be focused on D-All, but any differences in performance between the two datasets will be noted.

We conducted Tukey's Honestly Significant Differences (HSD) tests on the training, open-set, closed-set (training), closed-set (open) evaluation sets for D-All datasets with Gaussian and uniform timing of interaction. The corresponding plot is given in Appendix E.1.

[15] https://github.com/EMRResearch/ExtremeValueMachine

The results show that the proposed approaches (MMIBN and MMIBN:OL) decrease the long-term recognition performance loss significantly ($p < .001$) and substantially compared to FR, supporting the first part of our hypothesis **H1**. This finding is valid across all datasets (D-Ten and D-All for Gaussian and uniform times).

MMIBN performs equally well between Gaussian and uniform timing for D-All evaluation sets (i.e., no significant difference, but slightly worse in uniform), whereas, it does not perform at the same significance in D-Ten evaluation sets (performs significantly worse). MMIBN:OL performance changes depending on the dataset size and the evaluation set (performs equally well only in closed-sets in D-Ten, and for training and closed-set open in D-All). Nevertheless, the models have slightly or significantly higher loss in uniform timing as compared to Gaussian, supporting hypothesis **H4**.

Online learning does not perform better than MMIBN, because it increases the loss at all conditions. In fact, except for training set in D-All and D-Ten and closed-sets in D-Ten for uniform timing where MMIBN and MMIBN:OL perform at the same significance level, online learning is significantly worse, which is in contrast with our hypothesis **H2**.

Furthermore, the results show that soft biometric features (SB and SB:OL) are not able to identify a user on their own. In general, they perform significantly worse than FR. However, when the interaction is time patterned (Gaussian), SB performs better and closer to FR as compared to uniform timing. Especially for closed-set training in D-All, it is remarkable that SB features identify the user with the same significance level performance as FR. SB and SB:OL perform mostly equally well in D-All datasets, but SB:OL performs significantly worse in several evaluation sets in D-Ten.

EVM:FR performs significantly better ($p < .005$) than FR across all conditions. EVM:MM is significantly worse than EVM:FR ($p < .01$) and it does not perform better than FR in most conditions. This shows that although EVM is a good method for clustering face recognition data, it does not perform well with multi-modal data.

MMIBN significantly outperforms ($p < .001$) both EVM models across all conditions in both D-All and D-Ten. This proves that our proposed approach is significantly better than the state-of-the-art method for incremental open world recognition with multi-modal biometric information. However, EVM models use online learning instead of fixed learning rates, which could potentially lead to worse performance as observed for our model. Nevertheless, comparing EVM models to MMIBN:OL shows that MMIBN:OL significantly outperforms EVM models ($p < .05$ to $p < .001$) in most cases, except for uniform timing for open-set and closed-set (open) in D-All and open-set in D-Ten, in which, it performs equally well with EVM:FR.

MMIBN performs equally well between training and open-set cases as well as between closed-sets, which shows that the model scales well for an increase in users (from 100 to 200 users), suggesting that the proposed approach and the optimised weights can generalise. Similar to the results in [73], EVM performs equally well between those sets, showing that the change in model from batch updates to incremental updates have not changed its structure for scaling well. The models perform significantly better in closed-sets as compared to training or open-set due to the lack of unknown users in closed-sets (FAR= 0.0). Hence, loss only depends on DIR.

The models are trained on several examples of the users before the closed-set. The model performance improves with the increasing number of recognitions and stabilises towards the end (around 2000), as can be observed in Fig. 5. This supports our initial finding of performance difference between D-All and D-Ten, given that they perform equally well for the same number of recognitions. Initially, loss increases with increasing FAR, when the users are introduced to the system (represented by dots in the plot). As the number of recognitions increases, the introduction of a new user does not notably increase the loss as can be observed by the final three new users in the training set. Even though MMIBN models get better over time, they start performing consistently

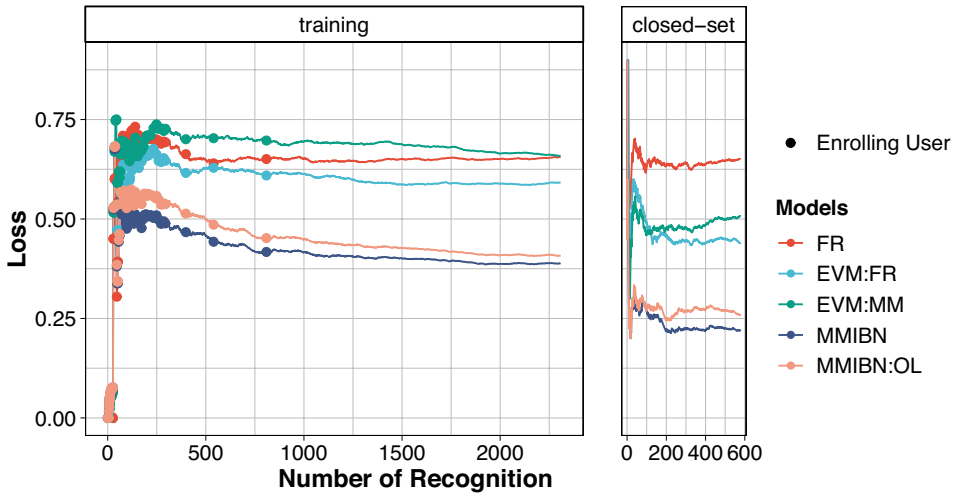


Fig. 5. The change of loss with the increasing number of recognitions for all samples dataset with Gaussian times (D-All$_{\text{Gaussian}}$) for training and closed-set (training). The loss decreases with the increasing number of recognitions.

better than both FR and EVM models throughout both training and closed-set after only a small number of recognitions ($15-48$ in training, $1-6$ in closed-set).

The sudden change at the beginning for the training set is due to the sequential calculation of loss for time plots: a previously enrolled person has not been identified correctly for the first time that changes DIR from 1.0 to 0.5 (one out of two enrolled users was incorrectly identified). Note that the introduction of new users is at random order due to ShuffledKFold function described in Section 6.2. The results for the open-set, as given in Appendix F, show a similar pattern of loss between open-set and closed-set (of the open-set cross-validation).

6.5.2 Open-Set Identification Metrics: DIR and FAR. The previously presented results confirm our claims that our proposed multi-modal Bayesian networks perform significantly better than FR, SB and EVM in long-term interactions. Nonetheless, analysing the open-set identification metrics allows us to understand how the models perform for enrolled and unknown users through DIR and FAR, respectively. The detailed presentation of Tukey's HSD results is shown in Appendix E.2.

The results show that the increase in DIR is significant ($p < .001$) and drastic, from 0.268 of FR to 0.657 with MMIBN and 0.561 with MMIBN:OL averaging over all the conditions in D-All (timing of interaction and evaluation set). That is a 38.9% increase in identifying the users correctly by using MMIBN, no matter the condition, which is more than double what FR is capable of providing. Hence, our hypothesis **H1** that the loss will be reduced and DIR will be increased using our proposed models as compared to FR alone is fully and strongly supported.

It should be noted that the increase in DIR provided by our network is significantly higher ($p < .001$) than DIR of soft biometrics (0.226 on average for Gaussian timing in D-All). This shows that soft biometric data are not sufficient to identify an individual, yet when combined with the primary biometric, they improve the identification rate significantly (38.9% in D-All, and 31.8% in D-Ten). This conclusion is supported by the datasets where the time of interaction is uniformly distributed (DIR of SB is 0.013 on average), that is, due to the high variability of time, the identification rate of SB is close to zero. Nevertheless, MMIBN performs equally well in Gaussian,

and uniform timing within all evaluation sets in D-All, and MMIBN:OL performs equally well in D-Ten. As previously noted in **H4**, the loss is (slightly or significantly) higher and DIR is (slightly or significantly) lower for all datasets and MMIBN models between Gaussian and uniform timing.

MMIBN significantly outperforms both EVM methods in DIR in all datasets ($p < .001$). EVM:FR has significantly higher DIR than FR and EVM:MM ($p < .001$). EVM:FR performs equally well between uniform and Gaussian timing in all datasets, because it is trained only on FR data. DIR of EVM:MM drops below that of FR for uniform timing for both D-All and D-Ten, which shows that EVM is not a model to be used with time information, since the pattern of interaction with the user might not be known beforehand. Similarly, MMIBN:OL provides worse performance for uniform timing in D-All, but it always performs significantly better than or equally well with EVM:FR.

FR performs similarly in open and closed-sets in terms of loss, because it has significantly low FAR compared to MMIBN models. While low FAR is a desirable feature, the underlying reason for low FAR is that FR has very poor recognition performance on larger datasets and fails to recognise the users, because the highest similarity score returned by the identifier is lower than the threshold ($\theta_{FR} = 0.4$). However, as described in Appendix C.2, this threshold ensures the lowest loss for FR.

FAR of the proposed models is high because of the combination of all modalities, which increase the probability of mixing the unknown user with an enrolled user. Possible solutions to this problem will be proposed in Section 7. For our proposed models, FAR in the training set is generally slightly less than that of open-set, because of the higher number of users enrolled, but there are no significant differences across the datasets for MMIBN, supporting that the model scales well to a larger dataset without a significant decrease in performance.

In the training set, there is no significant difference between MMIBN and EVM models for FAR, and MMIBN:OL performs significantly better than EVM models for uniform timing. In contrast to MMIBN, EVM provides significantly lower FAR in open-sets than in training sets. The authors state in [73] that this is due to its ability to tightly bound class hypotheses by their support.

6.5.3 User-Specific Analysis. Confusion matrices presented in Fig. 6 show how users were identified throughout the training set in D-All for a fold of the cross-validation, with 0 as the ID of the unknown user and the remaining numbers corresponding to IDs of the enrolled users. The heat map represents the percentage of identification of the user as the estimated user. Ideally, the diagonal should be all dark red if users are correctly identified. However, FR (item A) mostly identifies the users as unknown, resulting in the corresponding vertical axis of 0 to be mostly red and in a low FAR and a low DIR. MMIBN (item B) has mostly red coloured dots on the diagonal but has mixed users with other enrolled users as can be seen from light blue dots all over the matrix. MMIBN:OL shows a similar pattern with slight deviations.

Even though EVM:FR (item C) only uses FR information, its confusion matrix is different from that of FR. The misidentifications are highly concentrated on the final ten users, suggesting that either FR or EVM might be subject to the catastrophic forgetting problem. Using multi-modal data overcomes that problem, as can be seen for EVM:MM (item D) as misclassifications are evenly distributed, similar to MMIBN. However, the diagonals in EVM models have notably fewer reds than MMIBN.

The significant differences in identification of users over the 5-folds of cross-validation, as presented in Appendix E.3, shows another striking result. FR does not perform equally well amongst the users in that there are significant differences of identification. Our proposed approach MMIBN balances the performance amongst users, thereby, reducing any recognition bias in the system while improving the performance of the overall system significantly as compared to FR. Online learning (MMIBN:OL and EVM:FR) balances the performance further, in contrast to the decrease in performance compared to MMIBN. EVM:MM shows a similar pattern.

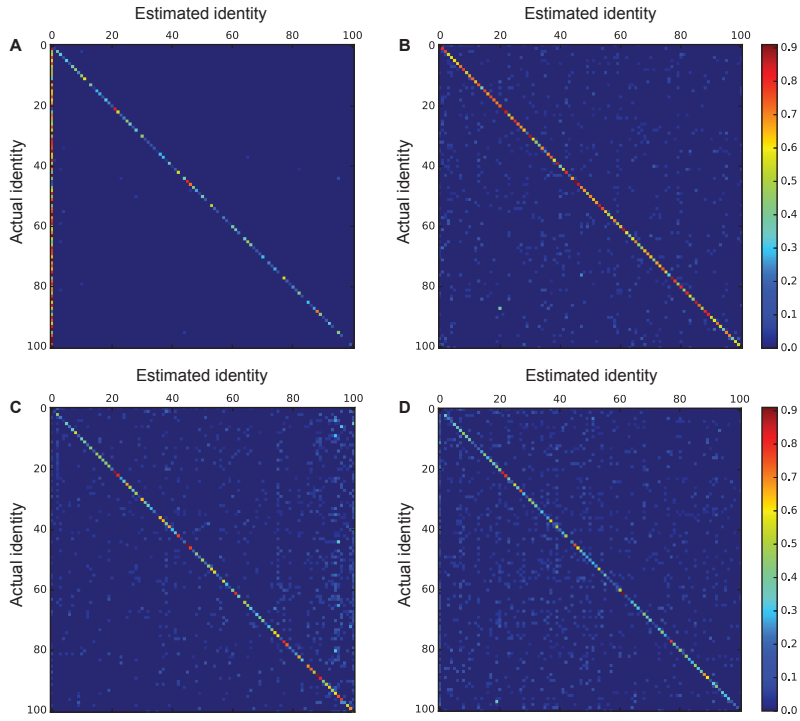


Fig. 6. Confusion matrices of user identification for second fold of cross-validation on D-All$_{\text{Gaussian}}$: **(A)** face recognition (FR), **(B)** proposed model (MMIBN), **(C)** incremental Extreme Value Machine (EVM) with FR data (EVM:FR), **(D)** incremental EVM with multi-modal data (EVM:MM). The heat map represents the percentage of identification as the estimated user. Ideally, if all users are correctly identified, the diagonal should be dark red, and the remaining of the matrix should be dark blue.

Fig. 7 demonstrates examples from D-All$_{\text{Gaussian}}$ where face recognition fails to recognise the user due to the low similarity score ($< \theta_{FR} = 0.4$), whereas, our proposed model identifies the user correctly based on soft biometric information. The quality of the estimation (Q) varies depending on the highest FR similarity score, as well as the disagreement between modalities. For example, for the third user (Sandra Oh), the highest FR similarity score (rank 1) is very low, corresponding to David Schwimmer who is 28 years old in the dataset, has a height of 185 with the enrolment time of interaction on Tuesday at 18:16. Age did not provide information to differentiate the user from the incorrect estimation, whereas, height and time of interaction increased the probability that the user is Sandra Oh, resulting in a correct estimation, but with a low quality score ($0.35 > \theta_Q = 0.013$). The second user (Gary Coleman) was identified correctly by FR with the highest similarity score close to, but slightly lower than θ_{FR}. This was enforced by the age estimation, and the time of interaction, which compensated for the incorrect recognitions of gender and height, to get a high quality score (7.44).

6.5.4 Real-Time Capabilities. In contrast to the state-of-the-art deep learning methods, the proposed models can run on a commercial robot with low computational power (on a single CPU of Pepper robot), and only require a small amount of time for execution. In addition to the time required from FR and other modalities (M= 0.14 s, SD= 0.001), MMIBN models take 0.01 second for recognition, significantly outperforming both EVM:FR and EVM:MM, which take 0.32 and 0.34,

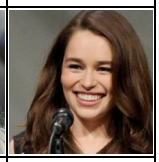
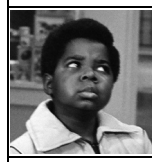
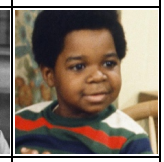
	Enrolment	*Probe:* 11		*Enrolment*	*Probe:* 2		*Enrolment*	*Probe:* 8	
									
	True Value	**Estimated Value**		**True Value**	**Estimated Value**		**True Value**	**Estimated Value**	
		FR	*BN*		*FR*	*BN*		*FR*	*BN*
ID	135	0 **[0.70]**	**135 [0.70]**	129	0 **[0.79]**	**129 [7.44]**	77	0 **[0.83]**	**77 [0.35]**
Name	Emilia Clarke	*FR (rank 1):* Angelina Jolie **[23.3%]**		Gary Coleman	*FR (rank 1):* Gary Coleman **[36.9%]**		Sandra Oh	*FR (rank 1):* David Schwimmer **[13.9%]**	
Gender	Female	Female **[72.7%]**		Male	Female **[88.1%]**		Female	Male **[66.3%]**	
Age	24	38 **[50%]**		10	7 **[100%]**		33	28 **[40%]**	
Height	157	152 **[8%]**		142	154.5 **[8%]**		168	172.7 **[8%]**	
Time	Saturday 10:15	Wednesday 17:35		Wednesday 13:41	Wednesday 13:53		Thursday 08:14	Thursday 07:57	

Fig. 7. Examples of true values and estimated values of modalities from our Multi-modal Long-Term User Recognition Dataset with Gaussian times (confidence values are given in brackets) using proposed non-adaptive multi-modal incremental Bayesian network with hybrid normalisation (referred to as BN in the figure). Highlights in red show the incorrect detection values. Face recognition was unable to recognise the users (0 represents unknown user) because the similarity scores were below the threshold (40%). Our proposed model is successful (highlighted in green) in correctly identifying the users with varying quality of estimations (shown in brackets underneath the ID) as a result of the information gathered from soft biometrics highlighted in blue. 8% confidence value of height corresponds to the $\sigma = 6.3$ cm in NAOqi.

respectively[16]. For enrolling new users, MMIBN requires a significantly lower amount of time (0.39 s, $p = .002$) for scaling the Bayesian network, compared to MMIBN:OL which takes 0.54 s, for which 0.17 s is due to online learning. There is no significant difference between MMIBN:OL and EVM models for enrolling (EVM:FR takes 0.48 and EVM:MM takes 0.52 s), with 0.20 and 0.23 s for online learning, respectively. The higher amount of time required for EVM:MM compared to EVM:FR shows that online learning takes longer time when there is more information to be learned per user. Note that the time required for MMIBN has decreased from 0.3 s in [43] to 0.01, as a result of optimising the MMIBN algorithm.

Moreover, in comparison to deep learning approaches, which require "big data" to be pretrained, our proposed models are able to start from a state of no enrolled users, learn users continuously and incrementally, and improve performance compared to FR after a small number of recognitions (e.g., 48 in Fig. 5).

[16]The results are given for D-All with Gaussian timing on the open-set.

7 DISCUSSION

Our findings showed that from our initial hypotheses **H1** and **H3** are fully supported, **H4** is supported for hybrid normalisation, and **H2** is rejected (Table 1). In this section, we will discuss the implications of our results, validate our assumptions, and offer other approaches for our models.

7.1 Dataset Size

In general, FAR and DIR is higher, and loss is lower in D-All than in D-Ten. The increase in DIR and the decrease in loss can be explained by the higher number of recognitions, which increases the performance over time. The increase in FAR can be due to different optimised weights for each dataset (see Fig. 19 in Appendix C.6). However, both datasets show similar patterns in differences between FR, SB and MMIBN models. Even though the number of samples per user is not the same in D-All, the fact that it performs equally well as D-Ten for the same number of recognitions shows that our equal priors assumption (Equation 4), which states that each user is equally likely to be seen, does not have any adverse effect on our proposed models.

While the weights of the biometric information differ based on the dataset size and the model, their positive values indicate that each modality is beneficial and effective in identifying users, and conditionally independent of each other, as supported by the learned structure (Appendix C.3). We suggest using the optimisation parameters (weights and quality threshold) that are optimised for D-All datasets since this dataset contains more samples. If the application is based on users appearing at specified times during a week (e.g., long-term therapy in a hospital), the optimised parameters for D-All$_{\text{Gaussian}}$ should be used; otherwise, it is better to use that of D-All$_{\text{Uniform}}$ (e.g., for companion robots). These optimised parameters generally perform significantly equivalent in both timing conditions in D-All for both models, as shown in Fig. 20, even though the timing of interaction does not provide enough information in the uniform timing case. Nonetheless, using different (or more accurate) identifiers for soft biometrics may result in a different set of weights and better recognition performance.

7.2 High False Alarm Rate

High FAR of the models is due to the trade-off between recognition and spotting unknown people, which is visible in Fig. 4. The value of α determines the importance of this trade-off in the loss function to ensure a higher number of correct recognitions in a long-term interaction. We found $\alpha = 0.9$ based on our assumption, that the average number of interactions is 10. Using a varying amount of samples (D-All) did not change the overall performance in terms of long-term recognition loss for the same number of total samples, when we compared D-All and D-Ten at the same amount of samples (800 for training and open-set and 200 for closed-sets). In Fig. 5, 71% of the users had less than 10 recognitions and 20% had more than 10, before the 800th recognition in D-All dataset. This finding shows that our choice of α did not negatively affect the results. Thus, instead of changing α for decreasing FAR, we would suggest using a variable threshold of quality (θ_Q) based on the number of users in the dataset to ensure that the quality is higher when the number of users is low.

The presented results are dependent on the noise level of the identifiers and the characteristics of the population (e.g., the distribution of parameters within the population). By using other algorithms for the identifiers or by setting a desired FAR depending on the application from Fig. 4, a different set of weights can be achieved with lower/higher FAR and consequently lower/higher DIR.

7.3 Online Learning

We initially assumed that all identifiers work equally well on all users based on the work in [45]. However, there can be changes in the person's appearance, the similarity between users, as well as

changes in time of interaction, which could negatively affect the visual identifiers and the time component of our models, respectively. We claimed that our online learning approach would adjust to these changes and perform better than the non-adaptive model (**H2**), but the second part of the hypothesis is not supported because online learning (MMIBN:OL) performed significantly worse or at the same significance as the non-adaptive MMIBN. The underlying reason might be the accumulating noise in the identifiers. We suggest three possible solutions for improving online learning: **(a)** identifiers with lower noise can be used, which can be difficult to achieve in real-world scenarios, **(b)** similar to the work in [20, 59], the learning rate η can be increased when there is a large error between the estimated parameter and its mean value, and decreased when convergence is reached, **(c)** confidence value of the identifiers or the quality of the estimation can be used to determine if the likelihoods should be updated at each iteration, to avoid updating when the noise is high. However, the average learned likelihoods in online learning showed that the initial parameter assumptions in Section 3.4 hold valid.

Online learning can also be applied to the weights of the MMIBN nodes to improve recognition performance over time based on the identifier accuracy, through decreasing or increasing the weights of the identifiers that are less or more accurate based on the data. We suggest applying online learning, similar to [40] or [56], on top of the optimised weights found in this work, which would allow adapting the MMIBN to work equally well (or better) with any (i.e., NAOqi or other) identifiers. However, a simpler approach is to apply Bayesian optimisation (of the weights and the quality of the estimation) on the Multi-modal Long-Term User Recognition Dataset, before deploying the MMIBN with other identifier algorithms to the real-world HRI applications.

FR does not perform equally well on users, as shown in Appendix E.3. Our proposed MMIBN models decrease the recognition bias in the system using multi-modal information. This finding is also confirmed for the uniform timing of interaction. Moreover, the first part of our hypothesis that online learning will adjust to these changes is supported, which allowed decreasing the bias of FR further. We can conclude that for long-term recognition our multi-modal incremental Bayesian networks not only perform better than FR alone in all datasets but also increases performance on each user to identify them equally well.

8 USER RECOGNITION IN LONG-TERM HUMAN-ROBOT INTERACTION IN THE REAL WORLD

8.1 User Identification Study

In our prior work [43], we proposed and applied a multi-modal weighted Bayesian network with online learning (MMIBN:OL) to a long-term HRI scenario (through the recognition architectures in Appendix A), where 14 participants (4 female, 10 male, of age range 24-40) interacted with the robot for 4 weeks in an office at the University of Plymouth (Fig. 8). The video showing the interaction for a known user is available online[17]. The study showed that our proposed approach enables and facilitates incremental identification in a real-world HRI scenario. Moreover, the optimised parameters on the real-world data showed an improvement (1.4% increase in DIR for closed-set and 4.4% in open-set) over face recognition (DIR= 0.903). Furthermore, MMIBN using minmax normalisation ($\text{MMIBN}_{\text{Minmax}}$) and MMIBN:OL with softmax ($\text{MMIBN:OL}_{\text{Softmax}}$) were the best performing methods on the data using zero weights on age and time of interaction. However, the resulting dataset was limited in terms of the number of participants and the characteristics of the participants, hence, the results and the optimisation parameters could not be generalised.

Correspondingly, we created the Multi-modal Long-Term User Recognition Dataset to optimise the parameters of our models and validate them on a large number of users in varying conditions

[17]Known user interaction: https://youtu.be/Ix98k6_-2Zc

Fig. 8. A user is interacting with the Pepper robot (SoftBank Robotics Europe) to confirm the identity that is estimated, during the user identification study in [43].

with a high variability of subject age and heights, which are highly challenging to obtain in an HRI experiment. The previous sections provided conclusive evidence that our proposed models are suitable for long-term user recognition, generalise well to new users and provide significantly more reliable identification than the state-of-the-art open world recognition model (Extreme Value Machine) and (NAOqi) face recognition alone. This section evaluates how the baselines and the optimised models in this work performed on the raw HRI data in comparison to the models in [43].

McNemar test is the best statistical method for comparing two classification algorithms that are run only once [28]. Cochran's Q test is an extension of the McNemar test for more than two groups. Thus, Cochran's Q test is applied to compare the identification of enrolled users (i.e., DIR) and new users (FAR) of all methods separately, and pairwise McNemar using Benjamini-Hochberg adjustment for multiple comparisons is applied as the post-hoc test [60]. The results show that there

Table 2. Pairwise McNemar test results on the identification of known users (DIR) for raw user identification data from [43]. Significant differences ($p < .05$) are highlighted in bold.

Model	FR	EVM:FR	EVM:MM	MMIBN Minmax	MMIBN:OL Softmax	MMIBN Uniform
EVM:FR	$\mathbf{p < .001}$ ($Z = -4.91$)	-	-	-	-	-
EVM:MM	$p = 0.06$ ($Z = 2.06$)	$\mathbf{p < .001}$ ($Z = 6.67$)	-	-	-	-
MMIBN Minmax	$\mathbf{p < .001}$ ($Z = -7.25$)	$p = .57$ ($Z = -0.742$)	$\mathbf{p < .001}$ ($Z = -6.99$)	-	-	-
MMIBN:OL Softmax	$\mathbf{p < .001}$ ($Z = -6.47$)	$p = .75$ ($Z = 0.404$)	$\mathbf{p < .001}$ ($Z = -6.15$)	$\mathbf{p = .04}$ ($Z = 2.29$)	-	-
MMIBN Uniform	$\mathbf{p < .001}$ ($Z = -6.63$)	$p = .76$ ($Z = -0.308$)	$\mathbf{p < .001}$ ($Z = -6.62$)	$p = .14$ ($Z = 1.63$)	$p = .23$ ($Z = -1.35$)	-
MMIBN:OL Uniform	$\mathbf{p < .001}$ ($Z = -5.74$)	$p = .64$ ($Z = 0.606$)	$\mathbf{p < .001}$ ($Z = -5.87$)	$\mathbf{p = .01}$ ($Z = 2.71$)	$p = .75$ ($Z = 0.365$)	$p = .07$ ($Z = 1.96$)

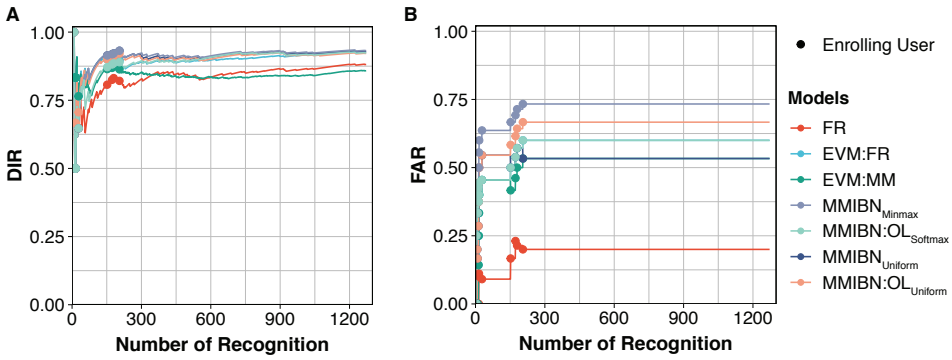


Fig. 9. Model performance on **(A)** known user identification (DIR) and **(B)** new user (incorrect) detection (FAR) on the raw data (1272 samples) from the user identification in HRI study [43] over 4-weeks for the optimised models in [43] (MMIBN$_{Minmax}$ and MMIBN:OL$_{Softmax}$), baselines (FR, EVM:FR, EVM:MM), and the optimised MMIBN models on the D-All$_{Uniform}$ dataset.

is a significant difference between all models ($p < .001$, $Q = 161.44$, df= 6) for DIR and the pairwise comparisons are shown in Table 2. MMIBN models with optimised parameters on the D-All$_{Uniform}$ dataset are used as the users were randomly encountered, however, no significant differences are observed between the models that were trained on the D-All$_{Gaussian}$ dataset (not shown for brevity). The results confirm that the optimised models in this work perform equally well as those that were optimised on the real-world data when the learning method is the same (e.g., comparing online learning models). Moreover, all MMIBN models significantly outperform FR (DIR= 0.881, L= 0.127) (supporting hypothesis **H1**) and EVM:MM (DIR= 0.858). Furthermore, the losses of MMIBN models are less than FR after only 39 recognitions. While the DIR of MMIBN$_{Minmax}$ is slightly higher (DIR= 0.932, L= 0.135) than MMIBN$_{Uniform}$ (DIR= 0.929, L= 0.117), MMIBN$_{Uniform}$ has the lowest loss. Similar to the previous results, online learning does not outperform the non-adaptive model, in contrast to our hypothesis **H2**. EVM:FR does not perform significantly different than the MMIBN models, however, it does not reach their performance over time (Fig. 9). Moreover, EVM models take substantially higher time to identify users (0.12 s for EVM:FR and 0.13 s for EVM:MM) than the MMIBN models (0.01 s for MMIBN and 0.03 s for MMIBN:OL). While there does not exist significant differences between the models in terms of FAR due to the low number of enrolled users, FR performs best (FAR= 0.2), followed by MMIBN$_{Uniform}$ and EVM:MM (FAR= 0.53).

8.2 Personalised Barista Robot

In a typical coffeehouse, baristas serve hundreds of customers per day and would not be able to recognise return customers or recall their preferences. A personalised robot could recognise a high number of customers, refer to them by name and recall and recommend their favourite orders, which could improve the customer experience and reduce the order time. In such an application, the customers will arrive sequentially at random times, and they need to be autonomously and incrementally added to the system with minimum time and effort from the customer. MMIBN corresponds to these requirements for incremental long-term user recognition in real-time. Consequently, the non-adaptive MMIBN with the optimised parameters on the D-All$_{Uniform}$ dataset was applied for identifying customers with a personalised barista robot (using the Adapted Pepper[18] robot) that recalls customer preferences [42].

[18]Created for MuMMER project: http://mummer-project.eu

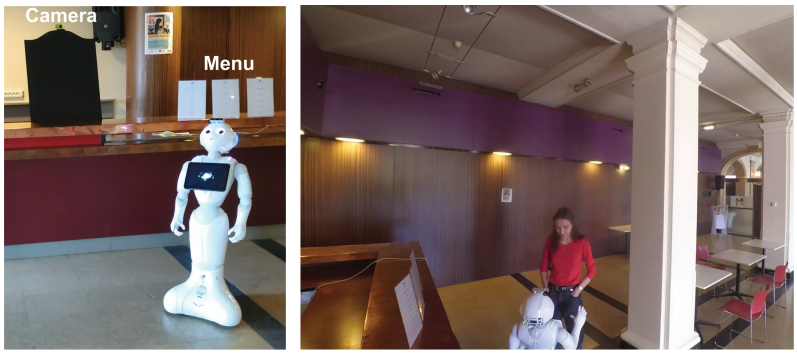


Fig. 10. Personalised barista robot [42] at the coffee bar of an international student campus, Cité Internationale Universitaire de Paris (France).

A 5-day HRI study with a generic (non-personalised) and a personalised barista robot was conducted in the coffee bar of an international student campus, Cité Internationale Universitaire de Paris (France), with 18 non-native English speakers (11 male, 7 female) within the age range of 22-47 (Fig. 10). Speech recognition was used to make the interaction more natural, and the confirmation of the estimated identity was implicitly taken through the dialogue (i.e., if the user does not oppose the estimated identity, the identity was assumed to be correct), in contrast to [43] where the user needed to explicitly confirm the identity through the tablet interface (as shown in Fig. 8). Also, ground truth values (gender, age, height, and an explicitly taken image) were not requested to reduce the effort required by the customer (i.e., step 9 in Fig. 15 was not used), thus, only the estimated values were used for enrolling users. However, users either did not realise that the estimated identity was incorrect or the identity was incorrectly confirmed due to speech recognition errors[19], which resulted in a high FAR (FAR= 0.786 for MMIBN, FAR= 0.286 for FR) and prevented some of the new users to be enrolled, showing the necessity of explicit user confirmation. Nonetheless, MMIBN performed better (DIR= 0.75, L= 0.304) than NAOqi FR (DIR= 0.5, L= 0.479 for 12 known user recognitions), supporting our hypothesis **H1**. Moreover, personalisation was found to mitigate the negative user experience, which suggests that user recognition plays an important role in long-term HRI. On average, 3.1 seconds (SD= 0.9) were taken to recognise users, which includes the time for user detection and the recognition module (Fig. 13) to obtain the biometric samples and the time for MMIBN to identify the user (0.01 s).

8.3 Personalised Socially Assistive Robot

Another area where personalisation can have an impact on long-term HRI is rehabilitation. Previous research shows that personalising the therapy improves user motivation and engagement, helps clinical staff in monitoring the progress of the patient, and facilitates rapport and trust over long-term interactions [19, 70, 75, 84]. Such improvements are desirable to improve adherence in cardiac rehabilitation, which is a long-term programme offered to those who suffered a cardiovascular event to accelerate recovery and reduce the risk of suffering recurrent events through structured exercise, education, and risk factor modification [35, 52]. Thus, in collaboration with medical specialists, a personalised socially assistive robot and a sensor interface [14, 15, 41, 54] were designed and deployed for long-term (18 weeks) cardiac rehabilitation programme at the Fundación Cardioinfantil-Instituto de Cardiología (Bogotá, Colombia), as shown in Fig. 11, for 5 months before the outbreak of

[19]Due to the errors in data, we could not apply statistical comparison between the MMIBN models and the baselines.

Fig. 11. Personalised socially assistive robot (using the NAO robot from SoftBank Robotics Europe) for long-term cardiac rehabilitation programme [14, 15, 41, 54] at the Fundación Cardioinfantil-Instituto de Cardiología (Bogotá, Colombia).

COVID-19 (which halted the programme at the clinic in March 2019). Because the robot is deployed in rehabilitation with non-expert users (e.g., doctors, nurses, patients), it should be autonomous and require minimal effort from users and medical staff [30]. Accordingly, an incremental user recognition system that does not need preliminary training is necessary for personalisation of the interaction, thus, MMIBN was chosen as the user recognition method. However, because the users will be generally encountered at patterned times (i.e., at their appointments twice per week), MMIBN with online learning with the optimised parameters on the D-All$_{\text{Gaussian}}$ dataset (MMIBN:OL$_{\text{Gaussian}}$) was used to evaluate its performance in a real-world interaction.

In contrast to the previous experiment [42], we used explicit confirmation of identity, in addition to the ground truth values for user enrolment, to avoid errors. The average recognition response time, which includes user detection, estimation of biometrics and identity, request of identity confirmation, the confirmation by the user on the tablet interface, and the updating of the model parameters (steps 1 to 10 in Fig. 15), was 24.8 seconds (SD= 15.5) for known user recognition, and 83.6 s (SD= 39.3) for new user enrolment, including the user to enter the ground truth values on the tablet (steps 1 to 18). Considering that the system is used by non-experts (patients), the time required is not substantial, especially because the patients take on average 9.39 s (SD= 17.46) to give a response to the tablet. MMIBN:OL took 0.04 s (SD= 0.01) for recognition.

Fig. 12 shows the performance of MMIBN:OL over time (with the increasing number of recognitions), and the performance of the other models on the real-world data is presented for comparison. 13 patients participated in the cardiac rehabilitation programme with the personalised robot, however, as observable from the figure, 30 enrolments were made to the system. The reason was a recurrent NAOqi face recognition failure that was never experienced in any of the prior studies, which resulted in erroneous user enrolments without registering the user's image to the face recognition database, thus, DIR dropped considerably. The experimenters at the hospital addressed the issue by re-enrolling some of the patients as new users, and the issue was resolved completely after the study by adding a threshold (e.g., 0.4) on NAOqi face recognition confidence. Nonetheless, Cochran's Q test shows significant differences between all models ($p < .001$, $Q = 21.49$, df= 4) for identifying enrolled users. Table 3 shows that there are significant differences between the MMIBN models and FR (DIR= 0.34, L=0.61), in addition to FR and EVM:MM. In contrast to our results in Section 6.5, MMIBN:OL performed slightly better than MMIBN in identifying known users (DIR= 0.38

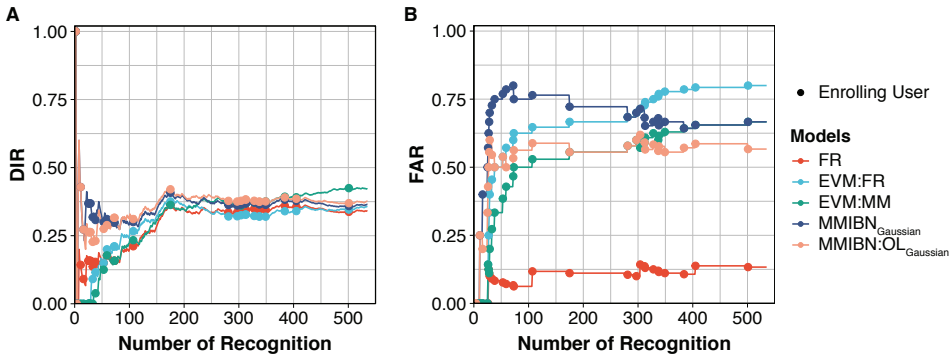


Fig. 12. Model performance on **(A)** known user identification (DIR) and **(B)** new user (incorrect) detection (FAR) throughout the cardiac rehabilitation programme with the personalised socially assistive robot, lasting 5 months (535 recognitions). MMIBN with online learning using optimised weights on the D-All$_{\text{Gaussian}}$ dataset (MMIBN:OL$_{\text{Gaussian}}$) was used for user identification during the programme. The performance of the other models on the real-world data are presented here for comparison.

for MMIBN:OL, DIR= 0.36 for MMIBN), notably better in identifying new users (FAR= 0.56 for MMIBN:OL, FAR= 0.67 for MMIBN), and achieved lower loss (L= 0.62 for MMIBN:OL, L= 0.64 for MMIBN), supporting our hypothesis **H2**, however, no significant differences are observed between the models. On the other hand, FR performed significantly better in FAR (FAR= 0.13, $p < .001$) than all baselines, because it identified most (63%) of known users as new. Because of the lower FAR and improving FR with re-enrolments, FR achieved a slightly lower loss than MMIBN:OL after 260 recognitions, thus, providing only partial support for our first hypothesis (**H1**).

While EVM:MM performs best overall in DIR and loss (DIR= 0.42, FAR=0.67, L=0.57), EVM:FR performs the worst of all models (DIR= 0.36, FAR= 0.8, L= 0.66), which is in contrast with the findings in Sections 6.5 and 8.1. Moreover, users were not recognised for the first 29 recognitions with EVM because of its tail size parameter that was optimised on the multi-modal dataset, and lowering it gave erroneous results. In contrast, only the first 4 estimations of MMIBN are discarded (i.e., users were identified as new, regardless of the model estimation), as in the multi-modal dataset. Furthermore, EVM models take 0.12 s for user recognition, which is substantially higher than MMIBN models (0.01 for non-adaptive model and 0.04 with online learning). These findings further

Table 3. Pairwise McNemar test results on the identification of known users (DIR) for the socially assistive robot study. Significant differences ($p < .05$) are highlighted in bold.

Model	FR	EVM:FR	EVM:MM	MMIBN$_{\text{Gaussian}}$
EVM:FR	$p = .44$ ($Z = -0.853$)	-	-	-
EVM:MM	$\mathbf{p < .001}$ ($Z = -4$)	$\mathbf{p = .02}$ ($Z = -2.8$)	-	-
MMIBN$_{\text{Gaussian}}$	$\mathbf{p = .02}$ ($Z = -2.56$)	$p = .68$ ($Z = -0.413$)	$\mathbf{p = .02}$ ($Z = 2.68$)	-
MMIBN:OL$_{\text{Gaussian}}$	$\mathbf{p = .02}$ ($Z = -2.65$)	$p = .44$ ($Z = -0.971$)	$p = .07$ ($Z = 2.01$)	$p = .44$ ($Z = -0.849$)

support that MMIBN models are the most reliable state-of-the-art open world user recognition method for HRI.

Overall, our findings on the Multi-modal Long-Term User Recognition Dataset and the real-world HRI experiments show that both of our proposed approaches perform better in recognising users than the state-of-the-art open world recognition method (Extreme Value Machine) and the NAOqi face recognition alone, supporting that our proposed user recognition models are suitable for incremental user identification in real-world HRI, and that they improve the recognition even when the identifiers are malfunctioning.

9 CONCLUSION

User identification is mostly regarded as a solved problem in the computer vision field. What remains unsolved is its application to the real world on low-computational power systems, such as commercial robots. The core problem that we face within HRI for personalising the interaction is to recognise unknown users and enrol them incrementally, which is classified as open world recognition. However, there exists a limited amount of research on this topic, and none of the available methods is evaluated on user identification. These methods use batch learning of classes instead of sequential learning, which is unlikely to be the case for HRI, because the users might not be available at the same time. In contrast, it is more likely that the same users will be encountered several times before the introduction of another.

Moreover, the computer vision field is not generally concerned with long-term interactions. Hence, correct identification of the enrolled users (DIR) and incorrect identification of the unknown users (FAR) are of equal value, whereas, the former is more valuable in long-term interactions since the same user is expected to be recognised several times, and the fraction of newly enrolled users will be much less. Furthermore, the appearance of the user may change over time, which requires updating the user database accordingly through online learning. In addition, combining soft biometrics, which are ancillary physical or behavioural characteristics (e.g., age) that can be extracted from primary biometric data (e.g., face) or available through other sources of information (e.g., time of interaction), can improve recognition accuracy.

In this work, we addressed these open challenges and presented a multi-modal incremental user recognition approach with online learning that is suitable for long-term HRI in the real world. We validated the approach within a variety of settings using an artificially generated multi-modal dataset, and through three real-world HRI experiments, thereby, extending the findings in our prior work [43] for a large number of users.

ACKNOWLEDGMENTS

This work has been supported by the EU H2020 Marie Skłodowska-Curie Actions Innovative Training Networks project APRIL (grant 674868), Royal Academy of Engineering IAPP project Human-Robot Interaction Strategies for Rehabilitation based on Socially Assistive Robotics (grant IAPP/1516/137), Colciencias (grant 813-2017), the Flemish Government (AI Research Program), the EU H2020 L2TOR project (grant 688014), and the EU FP7 DREAM project (grant 611391). The authors would like to thank the reviewers for their valuable suggestions for improving the presentation and analyses of the results, Valerio Biscione for his valuable suggestions in the design of the MMIBN, Pierre-Henri Wuillemin for his substantial help with the pyAgrum library, Ethan Rudd for sharing the Extreme Value Machine code, Jonathan Casas and Nathalia Céspedes Gómez for their help in integrating MMIBN within the personalised patient-robot interface for the cardiac rehabilitation programme, Mehdi Hellou for integrating MMIBN into the barista robot, the participants in our experiments for their time and efforts, and Hoang-Long Cao for his contribution of artwork in the human-robot interaction diagram (Fig. 1).

A RECOGNITION ARCHITECTURE

The recognition architectures presented in Fig. 13, 14, and 15 were used for the HRI experiments described in Section 8, namely, user identification in a research office [43], the personalised socially assistive robot for cardiac rehabilitation [41] and the personalised barista robot [42], as well as for evaluations on the Multi-modal Long-Term User Recognition Dataset (Section 6). The Recognition Module (Fig. 13) for NAOqi proprietary software was used to obtain the face similarity scores and gender, age and height estimations, along with the time of interaction, however, the last two parameters were artificially generated for the multi-modal dataset, as described in Section 5. The identifiers in the Recognition Module can be replaced with any software providing the same biometric estimations. The image, estimated and true identity, and ground truth values are automatically and incrementally fed into the system for the multi-modal dataset; in contrast, the image is taken (via the camera on the robot's or the tablet) when a user arrives, the estimated identity was announced to the user by a robot and confirmed by the user, and the ground truth values are entered by the user (through a tablet interface) in the HRI experiments. Fig. 16 illustrates user estimation and how the prior and likelihoods of the MMIBN change for incremental and online learning based on known and new users and the evidence from the identifiers.

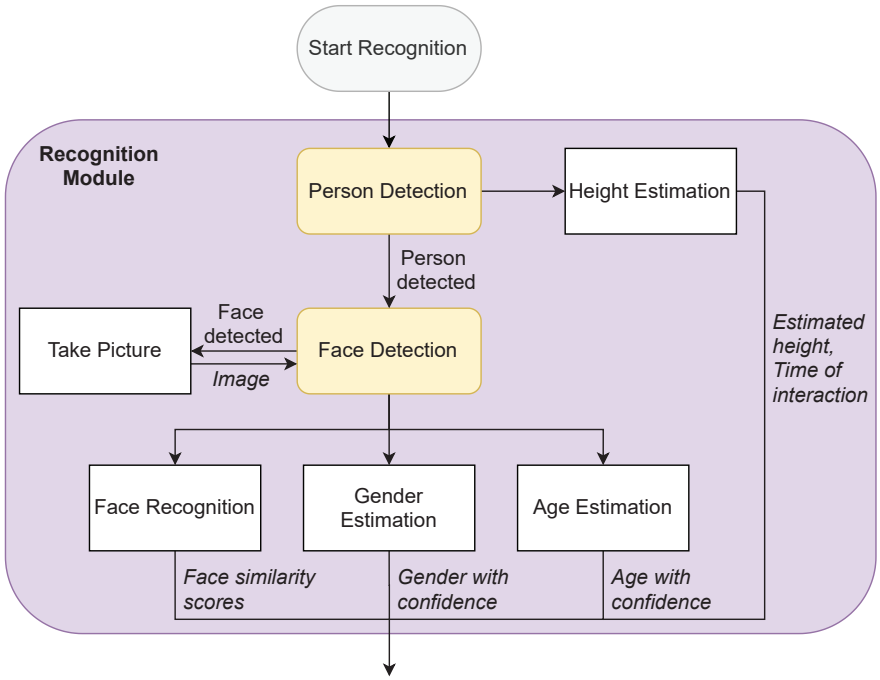


Fig. 13. Diagram of the recognition module. The yellow highlighted modules are proprietary software within NAOqi that are used to obtain the estimated modalities.

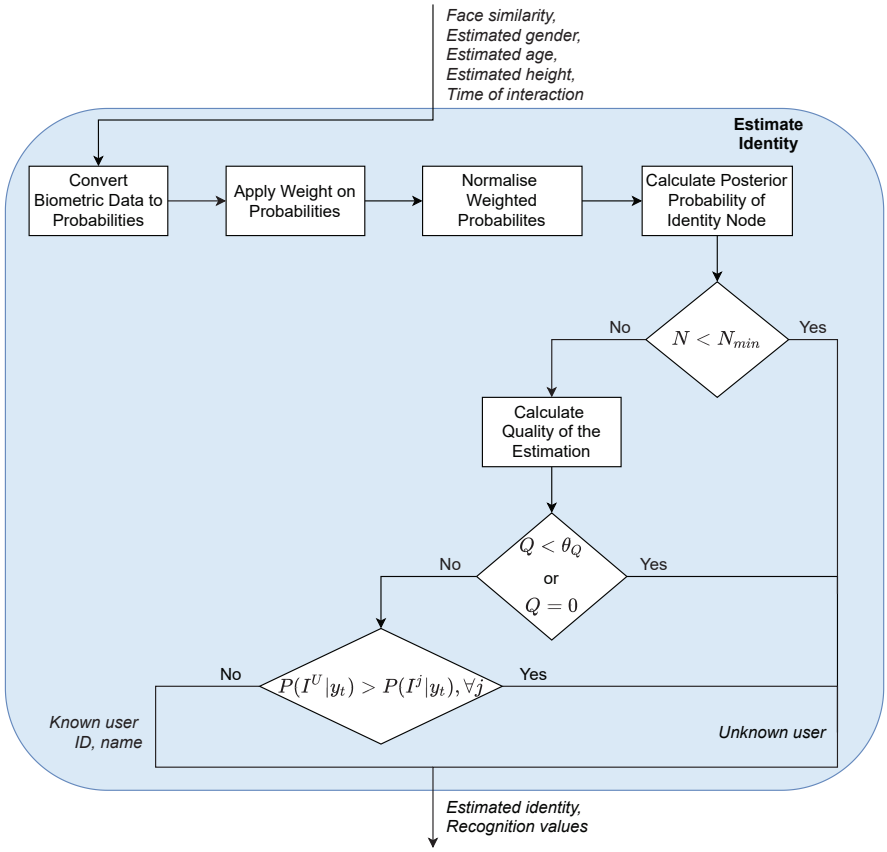


Fig. 14. Diagram of the estimation of the identity within the Multi-modal Incremental Bayesian Network (MMIBN). N is the number of the recognition, N_{min} is the minimum number of recognitions to ensure an identity is estimated correctly by MMIBN (taken as 5, as explained in Section 3.4), Q is the quality of the estimation in Equation 5) which is compared to the threshold (θ_Q), $P(I^j|y_t)$ is the posterior of the identity for user j and unknown user U in Equation 6.

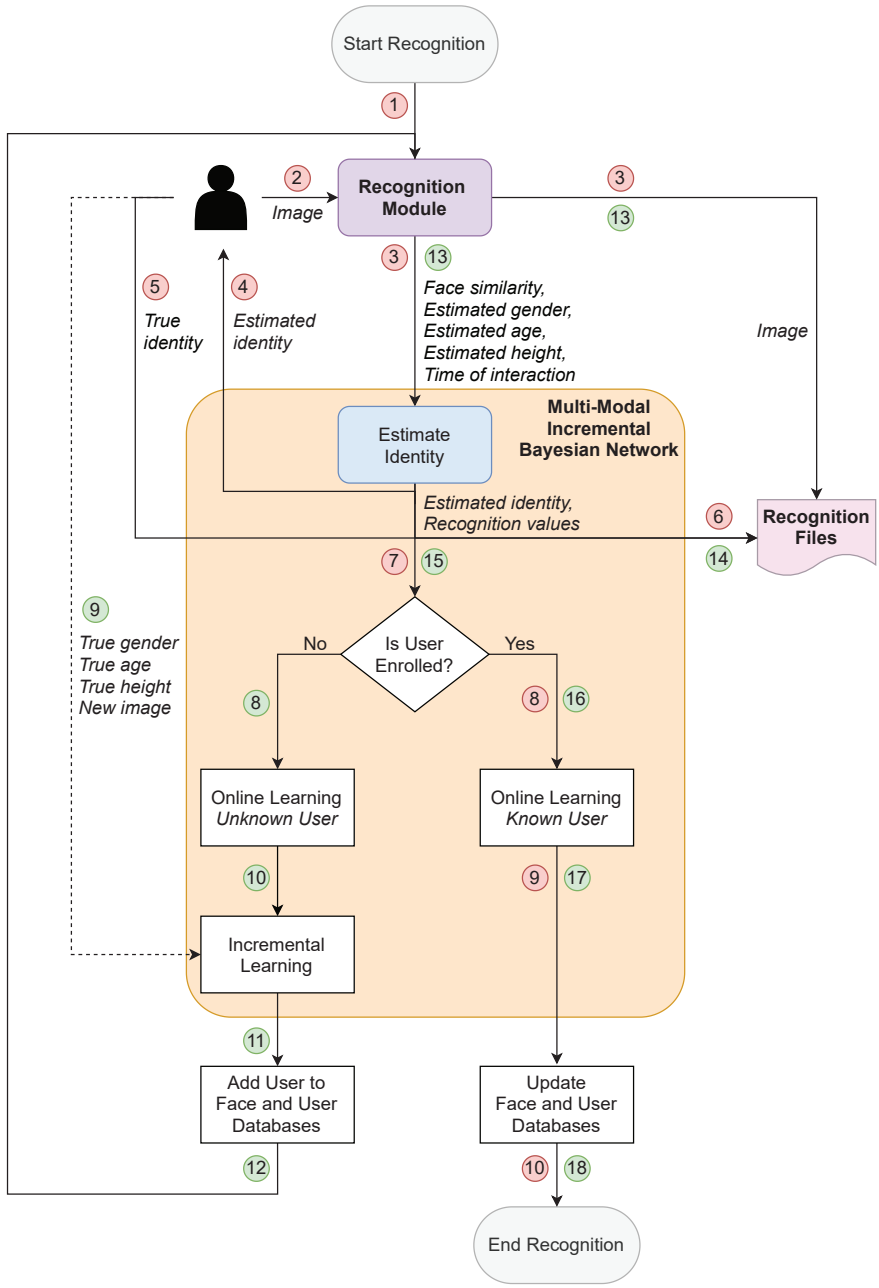


Fig. 15. Diagram of the recognition architecture for Multi-modal Incremental Bayesian Network with Online Learning (MMIBN:OL). Steps 8 and 16 are not used in non-adaptive MMIBN. The first 7 actions in the architecture are common to both known (enrolled) and new users. Actions 8-18 (in green) are performed for new users, whereas, 8-10 (in red) are performed for known users. Dashed line shows that ground truth values for gender, age and height are requested from the user and a new image is taken as input when the user is enrolling (if this step is skipped, estimated values will be used for enrolment). The recognition module and the estimation of the identity within the MMIBN are presented in Fig. 13 and 14, respectively.

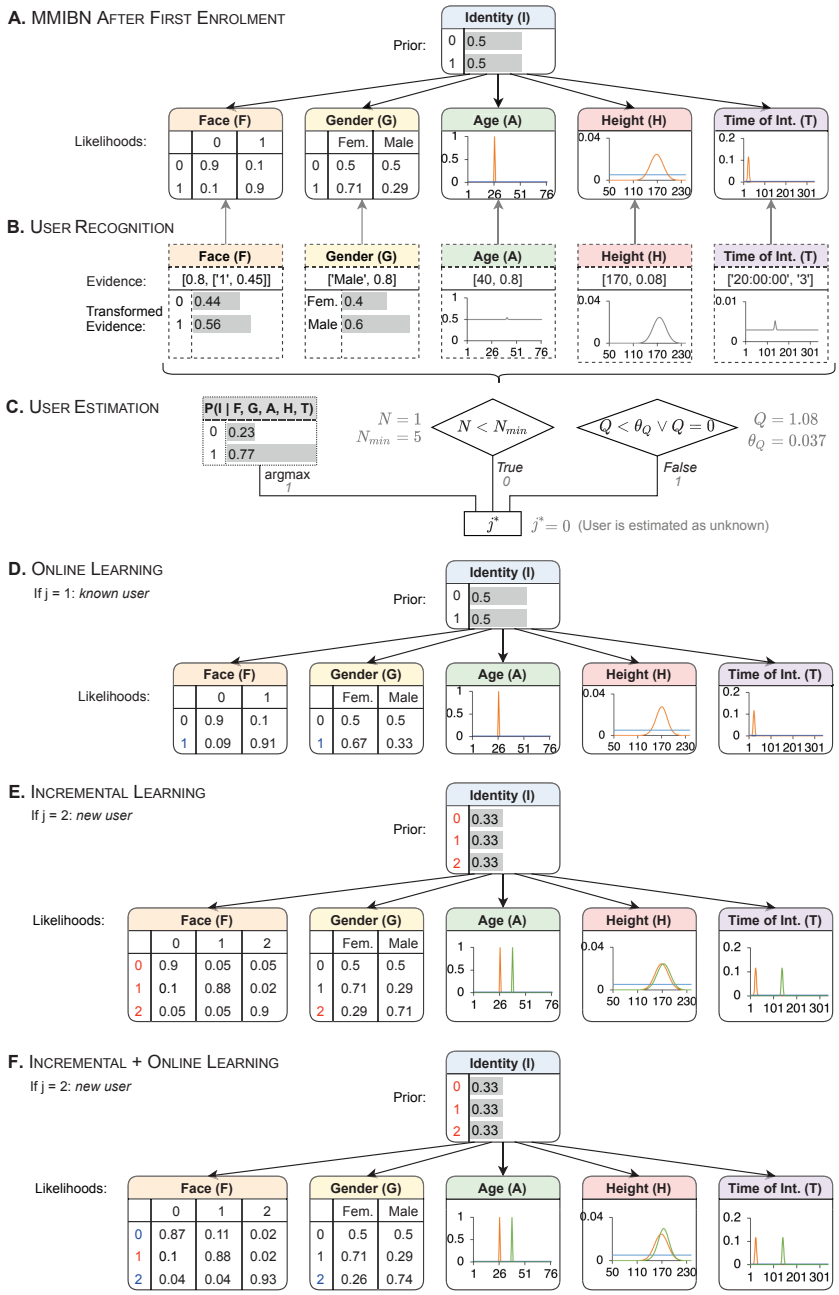


Fig. 16. Illustration of MMIBN: **(A)** initial model after the enrolment of the first user ($j = 1$), **(B)** user evidence from identifiers and transformed (weighted and normalised) evidence during user recognition, **(C)** estimated user (j^*) based on posterior score ($P(I|F, G, A, H, T)$), the minimum (N_{min}) number of recognitions (N), and quality of the estimation (Q) and its threshold (θ_Q), **(D)** model with online learning for a known user ($j = 1$), **(E)** model with incremental learning for a new user ($j = 2$), **(F)** model with incremental and online learning for a new user ($j = 2$). Non-adaptive model remains the same as in **(A)** for a known user ($j = 1$). Updated likelihoods for I, F and G are shown in red for incremental learning, and blue for online learning. For A, H and T nodes, $j = 0$ (unknown) is represented by light blue, $j = 1$ by orange and $j = 2$ by green.

B REPEATED K-FOLD CROSS-VALIDATION GENERATION

As described in Section 6.2, two methods are used to create the cross-validation repeats, as presented in Algorithm 1: **OrderedKFold**, where users are enrolled one after another and the enrolment order is different in each fold, and **ShuffledKFold**, where the user samples (probes) are shuffled, hence, users may be repeatedly seen before another user is enrolled.

Algorithm 1 Repeated K-Fold Cross-Validation Generation

function OrderedKFold(K, M) ▷ K is number of folds, M is the samples for each user in the dataset
  $k \leftarrow 1$
  while $k <= K$ **do** ▷ Create initial cross-validation set
    $SM \leftarrow$ shuffle order of M ▷ Enrollment order is different across each bin
    $B[k] \leftarrow SM[i][j : j+\text{length}(SM[i])/K]$ ▷ Divide user samples equally across each bin
    $V[k] \leftarrow$ stratified randomise order B ▷ Initial and final bins are different across K folds
    $k \leftarrow k + 1$
  return V ▷ Validation set
function ShuffledKFold(K, O) ▷ K is number of folds, O is the (previous) validation set
  $SP \leftarrow$ shuffle O ▷ Shuffle the order of the user samples in previous validation set
  $k \leftarrow 1$
  while $k <= K$ **do** ▷ Create initial cross-validation set
    $B[k] \leftarrow SP[j : j+\text{length}(SP)/K]$ ▷ Divide shuffled validation set across each bin
    $V[k] \leftarrow$ stratified randomise order B ▷ Initial and final bins are different across K folds
  return V ▷ Validation set
procedure RepeatedKFold(R, K, M) ▷ R is number of repeats, K is number of folds, M is the samples for each user in the dataset
  $C[1] \leftarrow$ OrderedKFold(K, M)
  $r \leftarrow 2$
  while $r <= R$ **do** ▷ Create cross-validation set for number of repeats
    $C[r] \leftarrow$ ShuffledKFold($K, C[r-1]$)
    $r \leftarrow r + 1$
  return C ▷ Repeated K-Fold Cross-Validation

C OPTIMISATION OF PARAMETERS

Initially, the loss parameter α and face recognition threshold is set as described in Sections C.1 and C.2. Furthermore, structural learning is applied to the data to validate the assumption of conditional independence in the Bayesian network (Section C.3). Subsequently, Bayesian optimisation is used to optimise the weights of the network and the threshold for the quality of the estimation (θ_Q). A total of 303 iterations is used for 5-fold cross-validation for each combination of the independent variables (for 40 conditions). The parameters are optimised by minimising the loss on the training set. By using the optimised parameters, 11 repeats of 5-fold cross-validation are conducted for each of the conditions to evaluate the effects of the independent variables on the open-set. For clarity of the presentation of results, we will initially analyse the results for 11-repeats of 5-fold cross-validation (in Section C.4), before presenting the optimised parameters from Bayesian optimisation. This would allow us to later analyse only the optimisation parameters (Section C.6) for the best performing normalisation method (Section C.5).

C.1 Loss Parameter

The loss parameter α (Equation 12) should be set to find the optimum FR threshold (θ_{FR}) and optimise the parameters in our network. As α increases, the fraction of correct recognitions of enrolled users (DIR) increases, but the fraction of the incorrect recognitions of unknown users (FAR) will increase. Based on our average number of observations assumption for long-term interaction ($\overline{n_o} = 10$), α becomes 0.9. For applications with fewer observations per user, α can be set accordingly.

C.2 Face Recognition Threshold

In FR, if the highest similarity score is below the face recognition threshold, θ_{FR}, the identity is classified as unknown. We examined how θ_{FR} influences the long-term recognition performance loss for the NAOqi FR in both D-Ten and D-All datasets, and noticed a decrease in performance (i.e., increase in loss) for $\theta_{FR} > 0.4$. Hence, we chose $\theta_{FR} = 0.4$ because it is the highest threshold giving the lowest loss to decrease FAR in our model, in agreement with our previous work in [43].

C.3 Bayesian Network Structure

In order to determine whether the conditional independence of the modalities (face, gender, age, height and time of interaction) given the identity of the user holds when the same input (i.e., image) is used to obtain multi-modal data, we applied structural learning of the Bayesian network on the Multi-modal Long-Term User Recognition Dataset using the pyAgrum [36] library. We used the all samples dataset with Gaussian times (D-All$_{\text{Gaussian}}$) with the identification estimations obtained from NAOqi proprietary algorithms (for face, gender and age estimations) and the artificially generated height estimations and time of interactions, as described in Section 5. Based on the requirements of the pyAgrum library, the multi-modal data is "simplified" by taking the best match evidence for modalities (i.e., confidence scores are not used) to allow structural learning. For instance, the most similar user (or unknown) is taken as the face recognition estimate by taking into account the face recognition threshold, and the evidence for gender, age and height are taken as the estimated values. Mandatory arcs (e.g., I -> F) between the identity node and the modalities are provided as prior structural knowledge, since the multi-modal information is used to determine the identity. Based on the Bayesian Dirichlet equivalent uniform (BDeu) score [39], all three methods available in the library (K2 algorithm [22], greedy hill-climbing search and local search with tabu list) found no other dependencies between the modalities, confirming the conditional independence.

We initially set the likelihoods to have much higher values for the true values, such as 0.9 for the face node (Equation 7) corresponding to the actual user and 0.99^{w_G} for the true gender. Average

learned likelihoods in online learning for 200 users confirm this assumption, with the mean for face node as 0.913 (SD=0.126), and the mean for the gender likelihood as 0.978 (SD=0.058).

C.4 Analysis of Variance of Independent Variables

Levene's test on the loss reveals ($F(10, 2189) = 0.026, p = 1$) that there is no significant difference in variances between the repeats, which indicates that our models are stable across repeats. ANOVA (Type-I) supports that there is no significant difference between repeats ($F(10, 2189) = 0.044, p = 1$), which shows that there is no significant difference between the ordered k-fold cross-validation and the shuffled k-fold, indicating that the model performs equally well for learning new users incrementally sequentially (similar to batch learning) and at random intervals (similar to a real-world scenario). Hence, we will only analyse the results of a single randomly selected repeat of 5-fold cross-validation. Since the model is stable across repeats, using a single repeat of cross-fold validation instead of independent test sets does not violate ANOVA assumption [7].

Due to the linear relation of loss with DIR and FAR in (12), there will be a correlation between the parameters. Pearson's product-moment partial correlation coefficient was computed to assess their relationships. The results show that there is a negative correlation between loss and FAR, $r(200) = -0.18, p = .009$, a positive correlation between loss and DIR, $r(200) = 0.99, p < .001$, but no significant correlation between FAR and DIR, $r(200) = 0.08, p = .25$.

A factorial ANOVA is conducted for analysing the primary and interaction effects of our independent variables. The results show that there are no significant primary effects for the model, $F(1, 160) = 1.50, p = .22$, and no significant interaction effects are found between the dataset size, timing of interaction, and model combination, $F(1, 160) = 0.01, p = .91$. Every other independent variable and their interactions are found to be significant ($p < .001$). This shows that the size of the dataset, timing of interaction and normalisation method have significant effects on the performance of the model, but online learning by itself does not provide significant improvement.

C.5 Normalisation Methods

A post-hoc analysis using Tukey's Honestly Significant Differences (HSD) test was conducted in which D-All and D-Ten datasets have been analysed separately for clarity, however, the results show similar patterns in both datasets. The corresponding Tukey's HSD test plots are presented in Fig. 17 and 18 (see Appendix E for the description of significance levels).

In both of D-All and D-Ten datasets, hybrid normalisation provides significantly lower loss ($p < .05$) in all conditions except for online learning in Gaussian timing for D-All ($p = .78$ in $\text{D-All}_{\text{Gaussian}}$), in which case it still provides the lowest mean for loss. Hence, our hypothesis **H3** is strongly supported, and the hybrid normalisation method is chosen for the remaining analyses.

While no significant differences are found in the primary effect of the learning method, there are significant differences between online learning and the non-adaptive model for hybrid normalisation. Online learning results in a higher loss for both datasets, which is in contrast with our hypothesis **H2**. The other methods do not show a stable pattern across conditions or datasets.

Most methods perform significantly worse in uniform timing of interaction (random interaction times), as compared to patterned interactions (Gaussian times), supporting our hypothesis **H4**. Softmax performs equally well on both models for D-All, but performs worse in uniform timing for D-Ten. Hybrid normalisation performs equally well for MMIBN in D-All but performs significantly worse in other conditions.

Hybrid normalisation performs better in all conditions and shows stability across varying conditions compared to the other methods. It achieves lower loss in D-All than in D-Ten, as a result of a higher number of samples in D-All (2280 in open-set) as compared to D-Ten (800 samples), which shows that the proposed model gets better with the increasing number of recognitions.

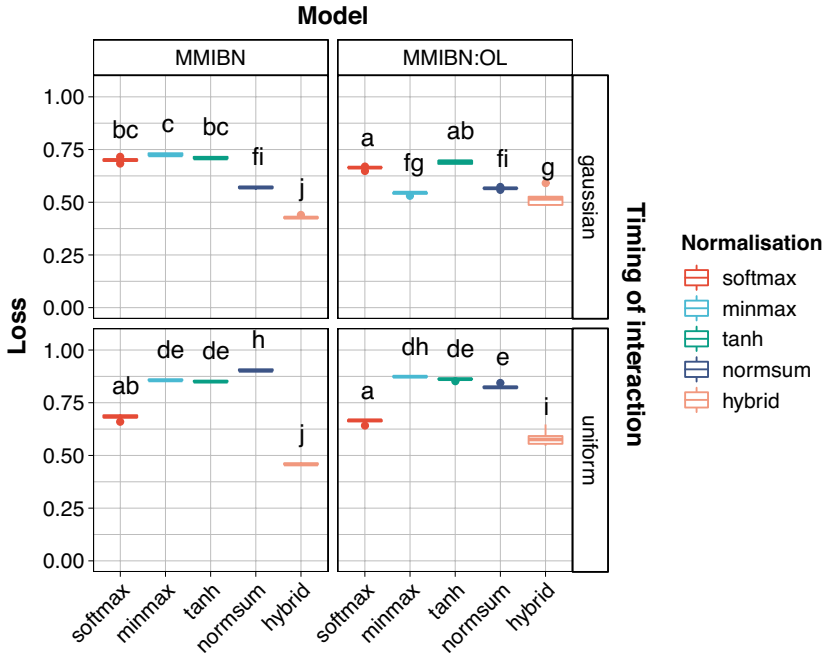


Fig. 17. Results of Tukey's HSD test of loss in the open-set for normalisation methods with optimised weights for all samples (D-All) dataset: softmax, minmax, tanh, normsum, and hybrid. Lower loss is better.

C.6 Weights and Quality of the Estimation

It seems to be self-evident that in the case of uniformly distributed time of interaction, online learning would provide worse results because the information provided by time will be unreliable. Hence, Bayesian optimisation should find a lower weight for the time parameter. The parameters corresponding to the optimum loss presented in Fig. 19, show otherwise. Weight for the uniform time is higher than that of the Gaussian for online learning in both datasets.

While the average standard deviation of NAOqi age estimation from the true age of the users[20] (i.e., the average standard deviation of error) is found to be 11.0 (which was 9.3 in [43]), age is found to be the most important parameter and height the least. This is in contrast with our findings in [43]. However, the results on the Multi-modal Long-Term User Recognition Dataset are more generalisable to larger populations, because of the higher number of users (200) and the diverse age range (10-63), in comparison to the limited number of users (14) and the narrow age range (24-40) in our prior work. Note that in the multi-modal dataset, we used the same standard deviation of height estimations (6.3 cm) as [43]. The standard deviation within age estimation (i.e., without using ground truths) is found to be 8.2, which is less than the standard deviation of error. NAOqi gender recognition rate[21] is found to be 0.9, and no difference is found between genders, that is, females and males are recognised equally accurately.

The optimised threshold for the quality of the estimation (θ_Q) is found to be less than 0.1 in each condition. The underlying reason is the disagreement of the modalities, which can decrease the

[20]The standard deviation of age estimation from the ground truth values are calculated per user, averaged over 200 users, and then averaged over 5-folds within the all samples dataset with Gaussian times (D-All$_{\text{Gaussian}}$).

[21]The gender recognition rate is the fraction of correctly estimated gender in the images (based on ground truths) of 200 users, averaged over 5-folds within the all samples dataset with Gaussian times (D-All$_{\text{Gaussian}}$).

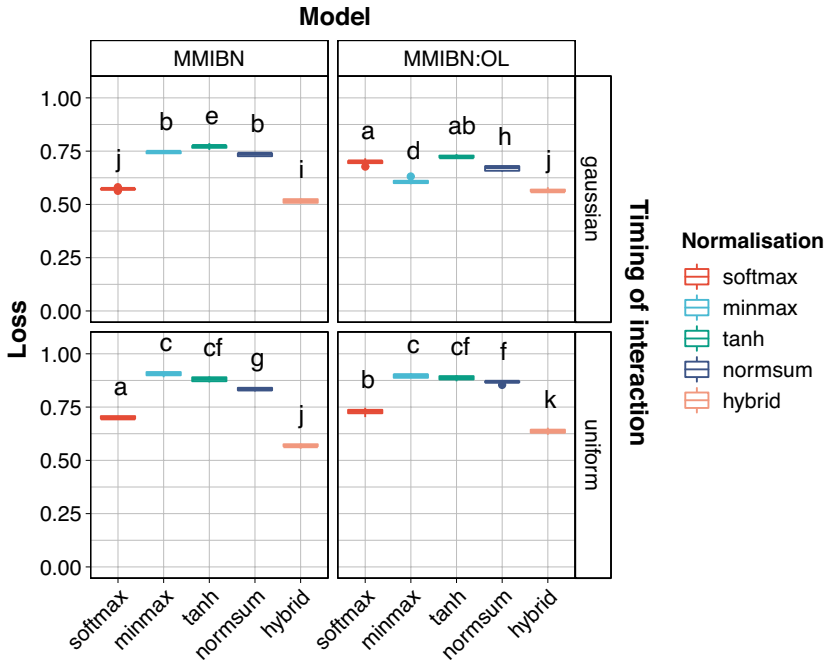


Fig. 18. Results of Tukey's HSD test of loss in the open-set for normalisation methods with optimised weights for ten samples (D-Ten) dataset: softmax, minmax, tanh, normsum, and hybrid. Lower loss is better.

differences in posterior probabilities because the results are combined through the product rule in the Bayesian network. When the modalities agree with high confidences (probabilities), the quality can be very high, such as $Q = 7.44$ as shown in Fig. 7 in Section 6.5 for the second user.

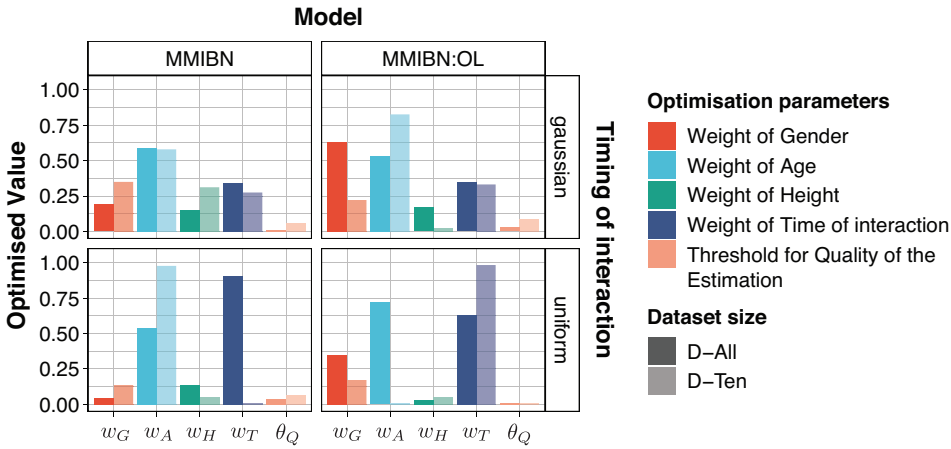


Fig. 19. Values of the optimised parameters (the weights of soft biometrics in the MMIBN models and the threshold for the quality of the estimation) through Bayesian optimisation of 303 iterations over 5-fold cross-validation for hybrid normalisation for all samples (D-All, represented with dark colours) and ten samples (D-Ten, represented with light colours) training sets.

D EXTREME VALUE MACHINE FOR INCREMENTAL ONLINE LEARNING

Extreme Value Machine[22] [73] (EVM) is a state-of-the-art open world classifier based on the Extreme Value Theory (EVT). However, it was only evaluated using batch learning, which is not suitable in a real-world HRI application, because the users will be encountered sequentially. Hence, we transformed the method for using sequential data and incremental online learning in order to compare the performance to our proposed methods[23].

The hyperparameters of EVM are tail size (τ, the number of points that constitute extrema for EVT), number of models to average (k), coverage threshold (ς, probabilistic threshold to designate redundancy between points), and open-set threshold (δ, if the maximum probability is below this threshold, the identity is estimated as unknown). The ranges considered for these hyperparameters in [73] are as follows: $100 - 32000$ for τ (can be minimum 2), $1 - 10$ for k, $[0.008, 0.186, 0.492, 1.0]$ for ς, and $[0.05, 0.1, ..., 0.3]$ for δ. Moreover, Euclidean distance or cosine similarity can be used as the distance function to compute margins for EVM.

As described in Section 3.4, we set MMIBN to declare the user as unknown in the first 4 recognitions, in order to allow the network to make meaningful estimations. This was achieved for EVM by setting $\tau = 3$. After the initial training, sequential learning is achieved by updating the model with a single data point (i.e., a single recognition) at each recognition, by setting $k = 1$. We optimised ς and δ over the ranges given, and found that $\varsigma = 1.0$ and $\delta = 0.05$ resulted in the lowest long-term recognition performance loss. Cosine similarity is used as the distance function, as it is stated in [73] that Euclidean distance led to poor performance for EVM.

It is important to note that in [73], $\tau = 33998$, $k = 6$, and $\varsigma = 0.5$. However, the authors stated that ς and k had a slight impact on performance (2% increase in accuracy and F1 score), whereas, the vast majority of performance variation was attributed to τ.

We use the same data with the structure described in Section 3.1 for evaluating MMIBN and EVM models. Note that for EVM models, the data is not normalised for face recognition similarity scores, and the normal curves for the remaining modalities are normalised through norm-sum (dividing by the total sum) because hybrid normalisation is a feature that we introduced in this paper for MMIBN and it is optimised for that structure. Using normalisation for face recognition does not change the performance of EVM:FR, but hybrid normalisation results in a poor performance for EVM:MM (DIR is 0.029).

[22] https://github.com/EMRResearch/ExtremeValueMachine

[23] Modified version of the Extreme Value Machine is provided online: https://github.com/birfan/MultimodalRecognitionDataset

E TUKEY'S HONESTLY SIGNIFICANT DIFFERENCES TEST PLOTS

In this manuscript, a letter representation is adopted for Tukey's HSD test plots. Levels that are not significantly different from each other at 0.95 confidence level ($p < .05$) are represented with the same letter over all the conditions, that is, each method is compared to all the other methods in different conditions. In other words, if two methods do not share a common letter, then there is a significant difference in performance between them. Multiple letters mean that the method is at the same significance level as multiple other methods.

E.1 Long-Term Recognition Performance Loss

Fig. 20 presents Tukey's HSD test results on the training, open-set, closed-set (training), closed-set (open) evaluation sets for D-All datasets with Gaussian and uniform timing of interaction. The results show that the proposed MMIBN model significantly outperforms FR and EVM in all of the datasets.

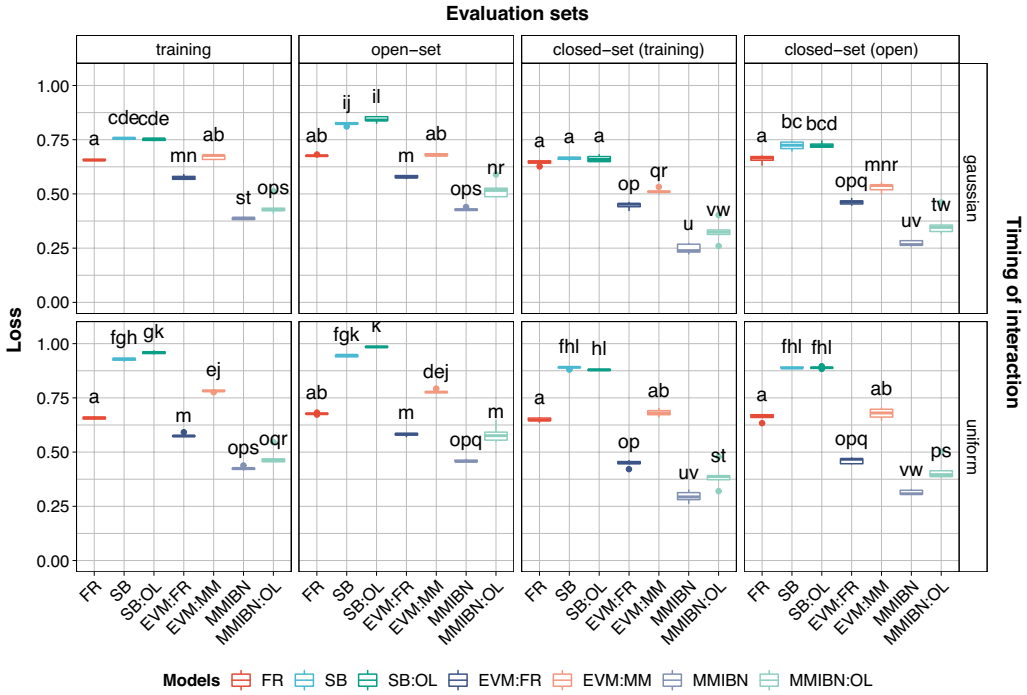


Fig. 20. Comparison of Tukey's HSD test results on loss for the proposed multi-modal incremental Bayesian network (MMIBN), face recognition (FR), soft biometrics (SB) with online learning condition (:OL), Extreme Value Machine with FR data (EVM:FR) and with multi-modal data (EVM:MM). The results are presented for training (100 users), open-set test (200 users), closed-set (training) (100 users) and closed-set (open) (200 users) for all samples dataset (D-All) for Gaussian and uniform timing of interaction. Lower loss is better.

E.2 Open-Set Identification Metrics: DIR and FAR

Tukey's HSD test results for DIR and FAR are presented in Fig. 21 and 22, respectively. The plot for DIR resembles highly that of Fig. 20 in a reversed direction, because of $\alpha * (1 - DIR)$ component of loss, whereby, $\alpha = 0.9$. DIR of MMIBN is significantly higher than FR and EVM in all datasets. The detailed analysis is presented in Section 6.5.2.

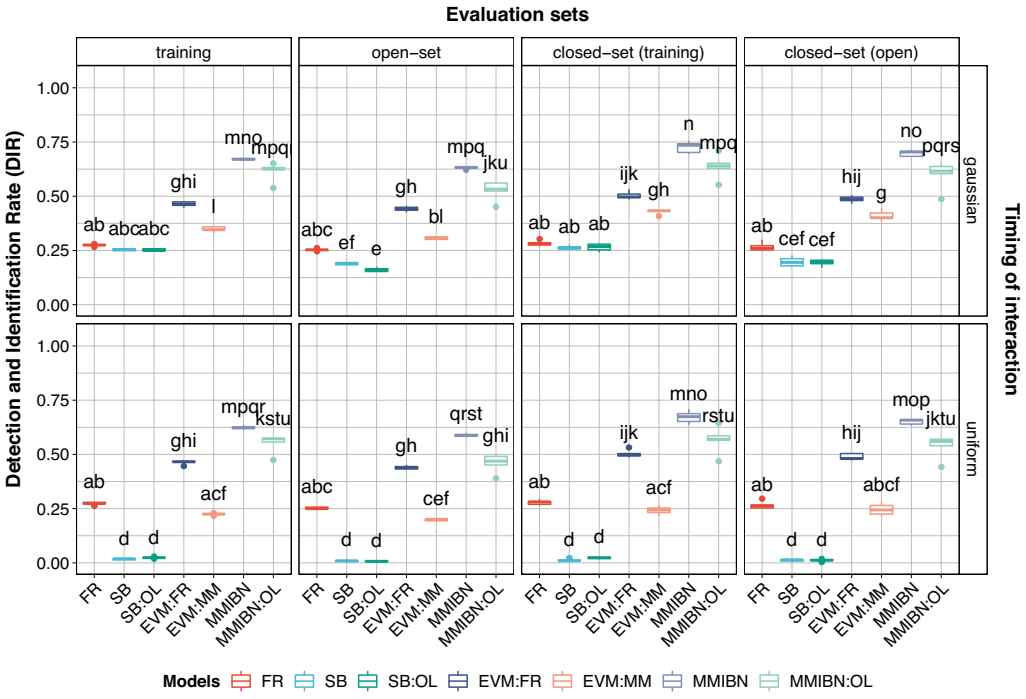


Fig. 21. Tukey's HSD test results for detection and identification rate (DIR) of all models for D-All datasets. Higher DIR is better.

FAR = 0 in closed-sets because all the users are previously enrolled. FR has a very low FAR in large datasets, because it predominantly identifies users as unknown. The combination of several modalities increases the probability to mistake a user for another user, which increase FAR in MMIBN.

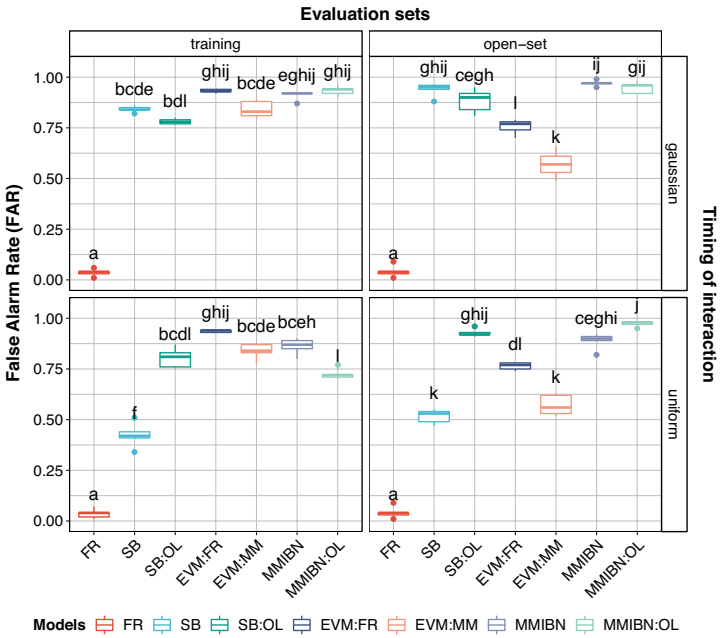


Fig. 22. Tukey's HSD test results for false alarm rate (FAR) of all models for D-All datasets. Lower FAR is better.

E.3 User-Specific Analysis

Fig. 23 presents the significant differences between the identification of users within the all samples dataset with patterned times. FR significantly performs better or worse for some of the users, whereas the combination of multi-modalities through our proposed model decreases the bias of FR. Online learning (EVM and MMIBN:OL) further mitigates the user recognition bias, in exchange for the performance, due to the accumulating noise in the identifiers.

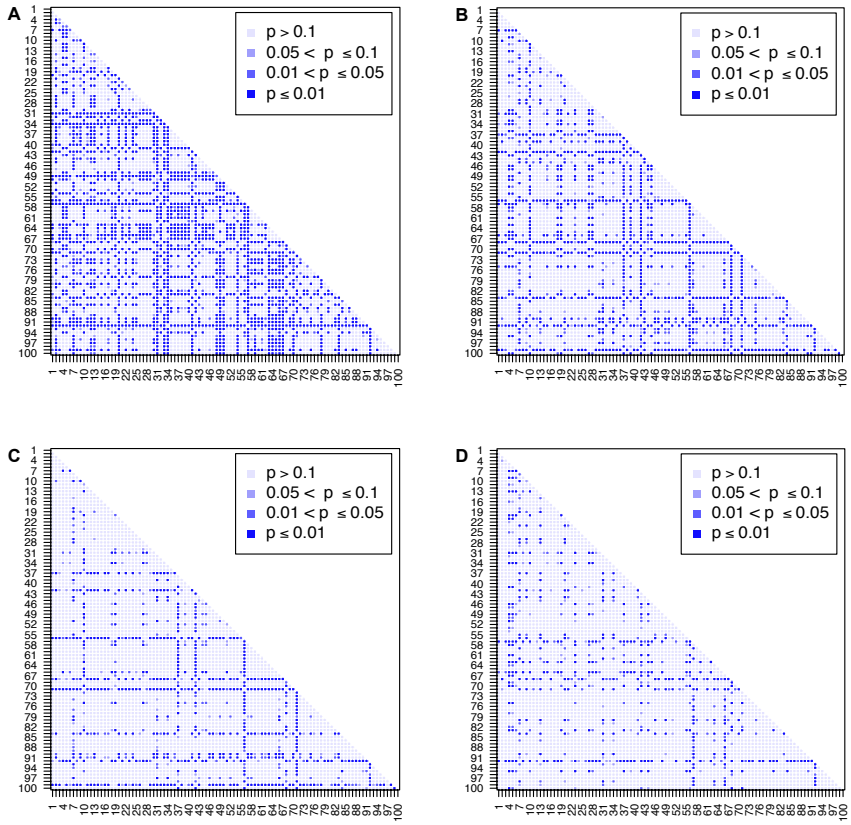


Fig. 23. Tukey's HSD test results for significant differences of user-based identification over 5-fold cross-validation on D-All$_{\text{Gaussian}}$: **(A)** face recognition (FR), **(B)** proposed model (MMIBN), **(C)** proposed model with online learning (MMIBN:OL), **(D)** Extreme Value Machine with face recognition data (EVM:FR). The darker blue colours represent significant differences, whereas lighter blue colours mean that the users are identified equally well.

F TIME PLOT FOR OPEN-SET RECOGNITION

The time plot for open-set recognition in Fig. 24 shows the change in long-term recognition loss with the increasing number of recognitions. The results are consistent with the results for the training set, presented in Section 6.5.1. MMIBN and MMIBN:OL have a higher loss in the open-set compared to the training, due to the higher number of users to recognise. EVM:FR has a lower loss during the enrolment period due to lower FAR compared to MMIBN models, and a higher DIR compared to EVM:MM, but the MMIBN models significantly outperform it overall and in the closed-set.

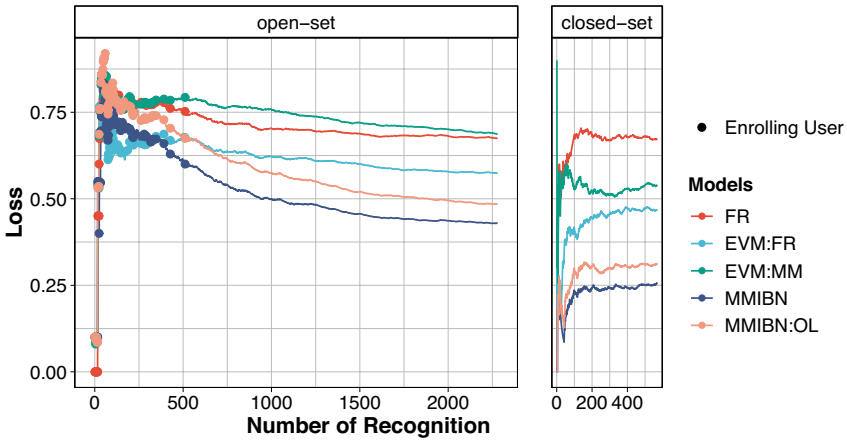


Fig. 24. The change of loss with increasing number of recognitions for all samples dataset with Gaussian times (D-All$_{\text{Gaussian}}$) for open-set and closed-set (open). The loss decreases with increasing number of recognitions.

REFERENCES

[1] Mohammad K. Al-Qaderi and Ahmad B. Rad. 2018. A Multi-Modal Person Recognition System for Social Robots. *Applied Sciences* 8, 3 (2018). https://doi.org/10.3390/app8030387

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications.* Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[3] Pierre Andry, Philippe Gaussier, Sorin Moga, Jean Paul Banquet, and Jacqueline Nadel. 2001. Learning and communication via imitation: an autonomous robot perspective. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 31, 5 (2001), 431–442. https://doi.org/10.1109/3468.952717

[4] Lijin Aryananda. 2001. Online and unsupervised face recognition for humanoid robot: toward relationship with people. In *IEEE-RAS International Conference on Humanoid Robots.* IEEE, Tokyo, Japan.

[5] Lijin Aryananda. 2009. Learning to recognize familiar faces in the real world. In *2009 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, Kobe, Japan, 1991–1996. https://doi.org/10.1109/ROBOT.2009.5152362

[6] Eric Bauer, Daphne Koller, and Yoram Singer. 1997. Update Rules for Parameter Estimation in Bayesian Networks. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence.* 3–13.

[7] Claudia Beleites and Reiner Salzer. 2008. Assessing and improving the stability of chemometric models in small sample size situations. *Analytical and Bioanalytical Chemistry* 390, 5 (2008), 1261–1271.

[8] Abhijit Bendale and Terrance Boult. 2015. Towards Open World Recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Boston, MA, USA, 1893–1902. https://doi.org/10.1109/CVPR.2015.7298799

[9] Abhijit Bendale and Terrance E. Boult. 2016. Towards Open Set Deep Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Las Vegas, NV, USA, 1563–1572. https://doi.org/10.1109/CVPR.2016.173

[10] E. S. Bigün, J. Bigün, B. Duc, and S. Fischer. 1997. Expert conciliation for multi modal person authentication systems by Bayesian statistics. In *Audio- and Video-based Biometric Person Authentication. Lecture Notes in Computer Science*, Vol. 1206. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0016008

[11] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc.

[12] Sofiane Boucenna, David Cohen, Andrew N Meltzoff, Philippe Gaussier, and Mohamed Chetouani. 2016. Robots Learn to Recognize Individuals from Imitative Encounters with People and Avatars. *Scientific Reports* 6 (2016), 19908. https://doi.org/10.1038/srep19908

[13] Sofiane Boucenna, Philippe Gaussier, Pierre Andry, and Laurence Hafemeister. 2014. A Robot Learns the Facial Expressions Recognition and Face/Non-face Discrimination Through an Imitation Game. *International Journal of Social Robotics* 6 (2014), 633–652. https://doi.org/10.1007/s12369-014-0245-z

[14] Jonathan Casas, Nathalia Céspedes Gomez, Emmanuel Senft, Bahar Irfan, Luisa F. Gutiérrez, Monica Rincón, Marcela Múnera, Tony Belpaeme, and Carlos A. Cifuentes. 2018. Architecture for a Social Assistive Robot in Cardiac Rehabilitation. In *2018 IEEE 2nd Colombian Conference on Robotics and Automation (CCRA).* 1–6. https://doi.org/10.1109/CCRA.2018.8588133

[15] Natalia Céspedes, Bahar Irfan, Emmanuel Senft, Carlos A. Cifuentes, Luisa F. Gutierrez, Mónica Rincon-Roncancio, Tony Belpaeme, and Marcela Múnera. 2021. A Socially Assistive Robot for Long-Term Cardiac Rehabilitation in the Real World. *Frontiers in Neurorobotics* 15 (2021), 21. https://doi.org/10.3389/fnbot.2021.633248

[16] Ching-Han Chen and Chia Te Chu. 2005. Fusion of Face and Iris Features for Multimodal Biometrics. In *Advances in Biometrics.* 571–580. https://doi.org/10.1007/11608288_76

[17] Tanzeem Choudhury, Brian Clarkson, Tony Jebara, and Alex Pentland. 1999. Multimodal person recognition using unconstrained audio and video. In *International Conference on Audio-and Video-Based Person Authentication.* 176–181.

[18] Nikhil Churamani, Paul Anton, Marc Brügger, Erik Flieundefinedwasser, Thomas Hummel, Julius Mayer, Waleed Mustafa, Hwei Geok Ng, Thi Linh Chi Nguyen, Quan Nguyen, and et al. 2017. The Impact of Personalisation on Human-Robot Interaction in Learning Scenarios. In *Proceedings of the 5th International Conference on Human Agent Interaction.* Association for Computing Machinery, 171–180. https://doi.org/10.1145/3125739.3125756

[19] Caitlyn Clabaugh, Kartik Mahajan, Shomik Jain, Roxanna Pakkar, David Becerra, Zhonghao Shi, Eric Deng, Rhianna Lee, Gisele Ragusa, and Maja Matarić. 2019. Long-Term Personalization of an In-Home Socially Assistive Robot for Children With Autism Spectrum Disorders. *Frontiers in Robotics and AI* 6 (2019), 110. https://doi.org/10.3389/frobt.2019.00110

[20] Ira Cohen, Alexandre Bronstein, and Fabio G. Cozman. 2001. *Online Learning of Bayesian Network Parameters.* Technical Report HPL-2001-55 (R.1). HP Laboratories.

[21] Ryan Connaughton, Kevin W. Bowyer, and Patrick J. Flynn. 2016. Fusion of Face and Iris Biometrics. In *Handbook of Iris Recognition.* Springer London, 397–415. https://doi.org/10.1007/978-1-4471-6784-6_17

[22] Gregory F. Cooper and Edward Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9 (1992), 309–347. https://doi.org/10.1007/BF00994110

[23] Claudia Cruz, L. Enrique Sucar, and Eduardo F. Morales. 2008. Real-time face recognition for human-robot interaction. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition.* IEEE, Amsterdam, Netherlands, 1–6.

[24] Antitza Dantcheva, Petros Elia, and Arun Ross. 2016. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security* 11, 3 (2016), 441–467. https://doi.org/10.1109/TIFS.2015.2480381

[25] Antitza Dantcheva, Carmelo Velardo, Angela D'Angelo, and Jean-Luc Dugelay. 2011. Bag of soft biometrics for person identification. *Multimed Tools Appl* 51 (2011), 739–777. https://doi.org/10.1007/s11042-010-0635-7

[26] Rocco De Rosa, Thomas Mensink, and Barbara Caputo. 2016. Online Open World Recognition. arXiv:cs.CV/1604.02275

[27] Rocco D. De Rosa, Francesco Orabona, and Nicolò Cesa-Bianchi. 2015. The ABACOC Algorithm: A Novel Approach for Nonparametric Classification of Data Streams. In *2015 IEEE International Conference on Data Mining*. IEEE, Atlantic City, NJ, USA, 733–738. https://doi.org/10.1109/ICDM.2015.43

[28] Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* 10, 7 (1998), 1895–1923. https://doi.org/10.1162/089976698300017197

[29] Geli Fei, Shuai Wang, and Bing Liu. 2016. Learning Cumulatively to Become More Knowledgeable. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, San Francisco, CA, 1565–1574. https://doi.org/10.1145/2939672.2939835

[30] David Feil-Seifer and Maja J. Matarić. 2005. Defining socially assistive robotics. In *9th International Conference on Rehabilitation Robotics (ICORR) 2005*. 465–468. https://doi.org/10.1109/ICORR.2005.1501143

[31] David Filliat. 2007. A visual bag of words method for interactive qualitative localization and mapping. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, Rome, Italy, 3921–3926. https://doi.org/10.1109/ROBOT.2007.364080

[32] Floris Gaisser, Maja Rudinac, Pieter P. Jonker, and David Tax. 2013. Online face recognition and learning for cognitive robots. In *2013 16th International Conference on Advanced Robotics (ICAR)*. IEEE, Montevideo, Uruguay, 1–9. https://doi.org/10.1109/ICAR.2013.6766498

[33] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. 2017. Generative OpenMax for Multi-Class Open Set Classification. arXiv:cs.CV/1707.07418

[34] A Gepperth and B Hammer. 2016. Incremental learning algorithms and applications. In *European Symposium on Artificial Neural Networks (ESANN)*. Bruges, Belgium.

[35] Catherine Giuliano, Belinda J. Parmenter, Michael K. Baker, Braden L. Mitchell, Andrew D. Williams, Katie Lyndon, Tarryn Mair, Andrew Maiorana, Neil A. Smart, and Itamar Levinger. 2017. Cardiac Rehabilitation for Patients With Coronary Artery Disease: A Practical Guide to Enhance Patient Outcomes Through Continuity of Care. *Clin Med Insights Cardiol.* 11 (2017). https://doi.org/10.1177/1179546817710028

[36] Christophe Gonzales, Lionel Torti, and Pierre-Henri Wuillemin. 2017. aGrUM: a Graphical Universal Model framework. In *International Conference on Industrial Engineering, Other Applications of Applied Intelligent Systems*. Springer, Arras, France. https://doi.org/10.1007/978-3-319-60045-1_20

[37] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

[38] Marc Hanheide, Sebastian Wrede, Christian Lang, and Gerhard Sagerer. 2008. Who am I talking with? A face memory for social robots. In *2008 IEEE International Conference on Robotics and Automation, ICRA*. IEEE, Pasadena, California, USA, 3660–3665. https://doi.org/10.1109/ROBOT.2008.4543772

[39] David Heckerman, Dan Geiger, and David M. Chickering. 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20 (1995), 197–243. https://doi.org/10.1023/A:1022623210503

[40] Carine Hue, Marc Boullé, and Vincent Lemaire. 2017. *Online Learning of a Weighted Selective Naive Bayes Classifier with Non-convex Optimization*. Springer International Publishing, Cham, 3–17. https://doi.org/10.1007/978-3-319-45763-5_1

[41] Bahar Irfan, Natalia Céspedes Gomez, Jonathan Casas, Emmanuel Senft, Luisa F. Gutiérrez, Monica Rincon-Roncancio, Marcela Munera, Tony Belpaeme, and Carlos A. Cifuentes. 2020. Using a Personalised Socially Assistive Robot for Cardiac Rehabilitation: A Long-Term Case Study. In *29th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 124–130. https://doi.org/10.1109/RO-MAN47096.2020.9223491

[42] Bahar Irfan, Mehdi Hellou, Alexandre Mazel, and Tony Belpaeme. 2020. Challenges of a Real-World HRI Study with Non-Native English Speakers: Can Personalisation Save the Day?. In *Companion of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM. https://doi.org/10.1145/3371382.3378278

[43] Bahar Irfan, Natalia Lyubova, Michael Garcia Ortiz, and Tony Belpaeme. 2018. Multi-modal Open-Set Person Identification in HRI. In *2018 ACM/IEEE International Conference on Human-Robot Interaction Social Robots in the Wild workshop*. ACM, Chicago, IL, USA.

[44] Anil Jain, Karthik Nandakumar, and Arun Ross. 2005. Score normalization in multimodal biometric systems. *Pattern Recognition* 38, 12 (2005), 2270–2285. https://doi.org/10.1016/j.patcog.2005.01.012

[45] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. 2004. Soft biometric traits for personal recognition systems. In *International Conference on Biometric Authentication (LNCS)*. Springer, Hong Kong, China, 731–738. https://doi.org/10.1007/978-3-540-25948-0_99

[46] Anil K. Jain and Unsang Park. 2009. Facial marks: Soft biometric for face recognition. In *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE Press, Cairo, Egypt, 37–40. https://doi.org/10.1109/ICIP.2009.5413921

[47] Anil K. Jain, Arun A. Ross, and Karthik Nandakumar. 2011. *Introduction to Biometrics*. Springer Publishing Company, Incorporated, Chapter Multibiometrics, 209–258.

[48] Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. 2014. Multi-class Open Set Recognition Using Probability of Inclusion. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Zurich, Switzerland, 393–409.

[49] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. 2002. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7 (2002), 881–892. https://doi.org/10.1109/TPAMI.2002.1017616

[50] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[51] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Chapter Parameter Estimation, 717–782.

[52] Willam E. Kraus and Steven J. Keteyian. 2007. *Cardiac Rehabilitation*. Humana Press, Totowa, NJ, USA. https://doi.org/10.1007/978-1-59745-452-0

[53] Alexis Lambert, Nahal Norouzi, Gerd Bruder, and Gregory Welch. 2020. A Systematic Review of Ten Years of Research on Human Interaction with Social Robots. *International Journal of Human–Computer Interaction* 36, 19 (2020), 1804–1817. https://doi.org/10.1080/10447318.2020.1801172

[54] Juan S. Lara, Jonathan Casas, Andres Aguirre, Marcela Munera, Monica Rincon-Roncancio, Bahar Irfan, Emmanuel Senft, Tony Belpaeme, and Carlos A. Cifuentes. 2017. Human-robot sensor interface for cardiac rehabilitation. In *2017 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, London, UK, 1013–1018. https://doi.org/10.1109/ICORR.2017.8009382

[55] Kuang-Chih Lee and D. Kriegman. 2005. Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, San Diego, CA, USA, 852–859. https://doi.org/10.1109/CVPR.2005.260

[56] Ming Li and Ren Zhang. 2020. Evolving a Weighted Bayesian Network for Consequence Assessment of Terrorist Attack. *IEEE Access* 8 (2020), 88282–88293. https://doi.org/10.1109/ACCESS.2020.2993016

[57] Sungsoo Lim and Sung-Bae Cho. 2006. Online Learning of Bayesian Network Parameters with Incomplete Data. In *Computational Intelligence*, De-Shuang Huang, Kang Li, and George William Irwin (Eds.). Springer Berlin Heidelberg, 309–314.

[58] Chao-Yu Lin, Kai-Tai Song, Yi-Wen Chen, Shuo-Cheng Chien, Sin-Horng Chen, Chen-Yu Chiang, Jyh-Her Yang, Yi-Chiao Wu, and Tzu-Jui Liu. 2012. User identification design by fusion of face recognition and speaker recognition. In *2012 12th International Conference on Control, Automation and Systems*. 1480–1485.

[59] Jinzhong Liu and Qin Liao. 2008. Online Learning of Bayesian Network Parameters. In *2008 Fourth International Conference on Natural Computation*, Vol. 3. IEEE, Jinan, China, 267–271. https://doi.org/10.1109/ICNC.2008.651

[60] Salvatore S. Mangiafico. 2016. *Summary and Analysis of Extension Program Evaluation in R*. Vol. version 1.18.8. rcompanion.org/handbook/

[61] Eric Martinson, Wallace Lawson, and Greg Trafton. 2013. Identifying people with soft-biometrics at fleet week. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction (HRI '13)*. IEEE Press, 49–56.

[62] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. 1995. Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights From the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review* 102, 3 (1995), 419–457.

[63] Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*, Gordon H. Bower (Ed.). Vol. 24. Academic Press, 109–165. https://doi.org/10.1016/S0079-7421(08)60536-8

[64] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. 2013. Distance-Based Image Classification: Generalizing to New Classes at Near-Zero Cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 11 (2013), 2624–2637. https://doi.org/10.1109/TPAMI.2013.83

[65] Simon Ouellet, François Grondin, Francis Leconte, and François Michaud. 2014. Multimodal biometric identification system for mobile robots combining human metrology to face recognition and speaker identification. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 323–328. https://doi.org/10.1109/ROMAN.2014.6926273

[66] German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2018. Continual Lifelong Learning with Neural Networks: A Review. arXiv:cs.LG/1802.07569

[67] Unsang Park and Anil K. Jain. 2010. Face Matching and Retrieval Using Soft Biometrics. *IEEE Transactions on Information Forensics and Security* 5, 3 (Sept 2010), 406–415. https://doi.org/10.1109/TIFS.2010.2049842

[68] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference*, Xianghua Xie, Mark W. Jones, and Gary K. L. Tam (Eds.). BMVA Press, Swansea, UK, 41.1–41.12. https://doi.org/10.5244/C.29.41

[69] P. Jonathon Phillips, Patrick Grother, and Ross Micheals. 2011. Evaluation methods in face recognition. In *Handbook of Face Recognition* (2nd ed.), Stan Z. Li and Anil K. Jain (Eds.). Springer Publishing Company, Incorporated, 553–556.

[70] Kathleen Richardson, Mark Coeckelbergh, Kutoma Wakunuma, Erik Billing, Tom Ziemke, Pablo Gomez, Bram Vanderborght, and Tony Belpaeme. 2018. Robot enhanced therapy for children with autism (DREAM): A social model of autism. *IEEE Technology and Society Magazine* 37, 1 (2018), 30–39.

[71] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. DEX: Deep EXpectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

[72] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision* 126, 2-4 (2018), 144–157.

[73] Ethan M Rudd, Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. 2018. The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence* 40, 3 (2018), 762–768.

[74] Debanjan Sadhya, Parth Pahariya, Rishi Yadav, Apoorv Rastogi, Ayush Kumar, Lakshya Sharma, and Sanjay K. Singh. 2017. BioSoft - a multimodal biometric database incorporating soft traits. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*. IEEE, New Delhi, India, 1–6. https://doi.org/10.1109/ISBA.2017.7947693

[75] Brian Scassellati, Laura Boccanfuso, Chien-Ming Huang, Marilena Mademtzi, Meiying Qin, Nicole Salomons, Pamela Ventola, and Frederick Shic. 2018. Improving social skills in children with ASD using a long-term, in-home social robot. *Science Robotics* 3, 21 (2018). https://doi.org/10.1126/scirobotics.aat7544

[76] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. 2013. Toward Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 7 (2013), 1757–1772. https://doi.org/10.1109/TPAMI.2012.256

[77] Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. 2014. Probability Models for Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 11 (2014), 2317–2324. https://doi.org/10.1109/TPAMI.2014.2321392

[78] W. J. Scheirer, N. Kumar, K. Ricanek, P. N. Belhumeur, and T. E. Boult. 2011. Fusing with context: A Bayesian approach to combining descriptive attributes. In *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, Washington, DC, USA, 1–8. https://doi.org/10.1109/IJCB.2011.6117490

[79] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2015). https://doi.org/10.1109/cvpr.2015.7298682

[80] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep Learning Face Representation from Predicting 10,000 Classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, OH, USA, 1891–1898. https://doi.org/10.1109/CVPR.2014.244

[81] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. IEEE Computer Society, Washington, DC, USA, 1701–1708. https://doi.org/10.1109/CVPR.2014.220

[82] Patrick Verlinde, Pascal Druyts, Gerard Cholet, and Marc Acheroy. 1999. Applying Bayes Based Classifiers for Decision Fusion in a Multi-modal Identity Verification System. In *Intl. Symposium. on Pattern Recognition*.

[83] Yunhong Wang, Tieniu Tan, and Anil K. Jain. 2003. Combining Face and Iris Biometrics for Identity Verification. In *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication* (Guildford, UK) *(AVBPA'03)*. Springer-Verlag, Berlin, Heidelberg, 805–813.

[84] Katie Winkle, Praminda Caleb-Solly, Ailie Turton, and Paul Bremner. 2018. Social Robots for Engagement in Rehabilitative Therapies: Design Implications from a Study with Therapists. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 289–297. https://doi.org/10.1145/3171221.3171273

[85] Waldemar Wójcik, Konrad Gromaszek, and Muhtar Junisbekov. 2016. Face recognition: Issues, methods and alternative applications. In *Face Recognition - Semisupervised Classification, Subspace Projection and Evaluation Methods*, S. Ramakrishnan (Ed.). InTech, Chapter 02. https://doi.org/10.5772/61471

[86] Rami Zewail, Ahmed Elsafi, Magdy Saeb, and Nourhan Hamdy. 2004. Soft and hard biometrics fusion for improved identity verification. In *47th Midwest Symposium on Circuits and Systems*, Vol. 1. IEEE, Hiroshima, Japan, I–225. https://doi.org/10.1109/MWSCAS.2004.1353967

[87] Zhijian Zhang, Rui Wang, Ke Pan, Stan Z. Li, and Peiren Zhang. 2007. Fusion of Near Infrared Face and Iris Biometrics. In *Advances in Biometrics*. Springer Berlin Heidelberg, 172–180. https://doi.org/10.1007/978-3-540-74549-5_19

[88] Yue Zhou and T. S. Huang. 2006. Weighted Bayesian Network for Visual Tracking. In *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 1. IEEE, Hong Kong, China, 523–526. https://doi.org/10.1109/ICPR.2006.1188